Department of Functional and Comparative Genomics

Institute of Integrative Biology

The effect of immunogenetic variability on human health: bioinformatics investigations from different perspectives

Louise Takeshita

Thesis submitted in accordance with the requirements of

The University of Liverpool for the degree of Doctor in Philosophy

October 2017

Abstract

The high level of diversity exhibited by genes coding for molecules involved in the immune system are critically involved with several aspects of human health. The ability of the immune system to recognize foreign pathogens or abnormal cells, while tolerating itself is achieved through a balance of various components and receptors in the immune system. Individual genotypes of immune genes such as HLA and KIR fine-tune this balance. Variability in those genes lead to diversity in the response to foreign molecules, and can also lead to intolerance to self-molecules in the form of autoimmune responses. Several past studies have found HLA and KIR polymorphisms to be associated with susceptibility or protection to a range of diseases and hypersensitivity to drugs. HLA has also a major role in transplantation, where transplanted tissues need to be as similar as possible to avoid rejection. The work described in this thesis contributes towards the knowledge of immunogenetic implications in associations with diseases, transplantation and adverse reactions to drugs using different bioinformatics approaches. First, it provides bioinformatics resources for a better understanding of the impact of immunogenetic diversity on human health in the form of two public databases, alongside with insights obtained through the analysis of their contents. The KIR and Disease Database (KDDB), described in Chapter 2, stores disease associations with KIR genes expressed in natural killer cells. The data within KDDB has been analysed to uncover trends within studies, in terms of the sets of KIR genes associated with susceptibility to, or protection from, auto-immune diseases, infectious diseases, pregnancy complications and cancer. The HLA Epitope Frequency Database (EpFreq-DB), described in Chapter 3, stores population frequencies of HLA epitopes, which are structural units on the surface of HLA molecules that have been increasingly associated with improvements in transplantation matching. An analysis has been performed to demonstrate global differences in the carriage of particular epitopes, and the potential functional consequences of using epitope mapping, instead of the more traditional allele matching for transplantation scenarios. Lastly, it investigates the molecular mechanisms underlying the association HLA polymorphisms with severe hypersensitivity caused by the anti-retroviral drug nevirapine, using molecular docking approaches. Developments described in this thesis contribute to better understanding of the influence of immune variability in human health and provides necessary knowledge to advances in personalized medicine.

Declaration

I hereby declare that the content of this thesis corresponds to my work, which has not been submitted for a degree at this University or any other institution. The contribution of others in the work presented in this thesis and other sources of information used in the text have been clearly acknowledged.

Chapter 2 shares much of its content with the web pages of the KDDB database (http://allelefrequencies.net/diseases/), which was developed by me (Chapter 2) and any contribution of others in the development of this database was clearly acknowledged. In Chapter 4, data generated by Prof. Munir Pirmohamed's group was utilised, and their results were the basis of the research question investigated in the chapter.

My contribution to the publications related to this research was as follows:

- **Takeshita LYC**, Gonzalez-Galarza FF, Dos Santos EJM, Maia MHT, Rahman MM, Zain SMS, Middleton D, Jones AR: **A database for curating the associations between killer cell immunoglobulin-like receptors and diseases in worldwide populations**. *Database* 2013, **2013**.
  Main author.
- **Takeshita LYC**, Jones AR, Gonzalez-Galarza FF, Middleton D: Allele frequencies **database**. *Transfus Med Hemotherapy* 2014, **41**:352–355.
  Main author.
- **González-Galarza FF**, Takeshita LYC, Santos EJM, Kempson F, Maia MHT, Da Silva ALS, Teles E Silva AL, Ghattaoraya GS, Alfirevic A, Jones AR, et al.: **Allele frequency net 2015 update: New features for HLA epitopes, KIR and disease and HLA adverse drug reaction associations**. *Nucleic Acids Res* 2015, **43**:D784–D788.
  Shared first authorship.
- Carr DF, Bourgeois S, Chaponda M, **Takeshita LY**, Morris AP, Cornejo Castro EM, Alfirevic A, Jones AR, Rigden DJ, Haldenby S, et al.: **Genome-wide association study of nevirapine hypersensitivity in a sub-Saharan African HIV-infected population**. *J Antimicrob Chemother* 2017, **72**:1152–1162.
  Co-author with molecular docking analysis.

Louise Takeshita

Acknowledgments

I would like to thank my primary supervisor Prof Andy Jones for guiding me in this research journey. Andy has been an inspiring supervisor and has provided me outstanding support during the PhD, encouraging me to develop and follow my own ideas, being understanding at times of hardship, meeting me regularly for updates and giving me advice on improving work efficiency. I am truly grateful for what I have learned under his supervision, not only as a researcher, but as an individual. I would also like to thank my co-supervisor Prof Derek Middleton for the support with immunogenetics knowledge from a practical perspective and for facilitating collaborations with other researchers.

I am grateful to Prof Sir Munir Pirmohamed and Dr Dan Carr for the opportunity of collaborating with their research, and to Dr Dan Rigden for helping me get started with molecular docking.

I am also grateful to Dr Eduardo Santos for encouraging me to apply for this PhD program, and to CNPq Brazil for funding my research.

I am deeply grateful to all my office colleagues for the company and support, especially Dr Faviel Gonzalez for proofreading texts, Dr Tony McCabe for the informatics support, and Dr Achchuthan Shanmugasundram for the fun times doing module assessments.

I would also like to extend my thanks to Prof Luciane Mello, Stefany, Gabi and Carina, for always being there when I needed.

Finally, my biggest thanks to my parents and my husband Osmar, for their constant love, understanding and support.

"Valeu a pena? Tudo vale a pena, se a alma não é pequena. "

"Was it worth it? Everything is worthy. If the soul is not small."

Fernando Pessoa

Table of Contents

Abstract i

Declaration ii

Acknowledgments iii

Table of Contents v

List of Figures ix

List of Tables xiii

Chapter 1 1

Investigations in Immunogenetics 1

1.1 The human immune system 1

1.1.1 Innate Immune System 2

1.1.2 Adaptive Immune System 3

1.1.3 Immunoglobulins: Humoral effectors of the adaptive immune system 5

1.1.4 Natural Killer Cells: Lymphocytes of the innate immune system 8

1.2 Genetics of the immune system 11

1.2.1 Population Genetics 12

Hardy-Weinberg Equilibrium 13

Linkage Disequilibrium 14

Haplotype Frequency Estimation 14

1.3 The Major Histocompatibility Complex (MHC) 15

1.3.1 Human Leucocyte Antigens (HLA) 16

HLA nomenclature and resolution levels 19

HLA evolution and diversity 20

HLA association with human pathologies 21

1.4 Killer-cell Immunoglobulin-like Receptors (KIR) 22

KIR associations with human pathologies 26

1.5 HLA matching in transplantation 27
1.5.1 Mechanisms of transplant rejection 27
1.5.2 HLA matching for renal transplants 28
1.6 Immunogenetic databases 30
1.6.1 IMGT/HLA and the Immuno Polymorphism Database (IPD) 30
1.6.2 Allele Frequency Net Database (AFND) 31
1.7 Research Aims 32
Chapter 2 33
A database for curating the associations between killer-cell immunoglobulin-like receptors and diseases in worldwide populations 33
2.1 Abstract 33
2.2 Introduction 34
2.3 Methods 36
2.3.1 Data curation 36
2.3.2 Implementation 38
2.3.3 KDDB Data Analysis 39
2.4 Results 41
2.4.1 Website organization 41
2.4.2 KDDB Content and Geographical Distribution 44
2.4.3 Descriptive analysis of KDDB data 48
2.5 Discussion 55
2.6 Conclusions 58
Chapter 3 60
HLA epitope matching in world populations 60
3.1 Abstract 60
3.2 Introduction 61
3.2.1 Anti-HLA Alloantibody Detection 62
3.2.2 HLA Epitopes and Structural Matching 63

3.3 Methods 68
3.3.1 HLA Epitope Definitions 68
3.3.2 HLA Population Data 69
3.3.3 Epitope Frequency Calculation 69
3.3.4 EpFreq-DB Implementation 72
3.3.5 HLA epitope alloreactivity analysis 73
HLA epitope profiles from populations 73
Identification of HLA epitopes specific to anti-HLA alloantibodies 75
Minimum HLA epitope sets 76
HLA allele and epitope 'negative' matching in populations 77
3.4 Results and Discussion 77
3.4.1 Validation of HLA epitope frequencies estimated from HLA haplotype frequencies 77
1.1.1 Website organization 78
3.4.2 Worldwide HLA epitope frequency distribution 81
3.4.3 HLA allele and epitope matching of alloreactivity profiles 83
3.5 Conclusion 92
Chapter 4 94
Structural basis of the association of HLA-C*04:01 with nevirapine adverse drug reactions 94
4.1 Abstract 94
4.2 Introduction 95
4.3 Methods 99
4.3.1 HLA-C*04:01 crystal structure 99
4.3.2 Choice of control HLA molecules 99
4.3.3 Modelling of control HLA molecules 100
4.3.4 Nevirapine Docking 101
4.4 Results and Discussion 103

4.4.1 Putative drug interaction with position 9 and 99 (NVP2) 112
4.4.2 Putative drug interaction with position 14 115
4.4.3 Putative drug interaction with residue Phe116 and Arg156 116
4.4.4 Critical evaluation of using docking tools for ADRs associated with HLA 118
4.5 Conclusion 118
Chapter 5 120
Discussion, general conclusions and future work 120
Appendix A 123
Appendix B 153
Appendix C 162
Bibliography 168

List of Figures

Figure 1.1: Components of the human immune system and the integration of innate and adaptive systems 2

Figure 1.2: Immunoglobulin structure from different perspectives 6

Figure 1.3: Complementary Determining Regions (CDR) loops in an antigen-binding site 7

Figure 1.4: Mechanisms determining NK cell activation according to an updated version of the "missing-self" 10

Figure 1.5: Location of MHC in chromosome 6 and genetic organization of class I, class II and class III regions 16

Figure 1.6: Structure of the HLA Class I and Class II molecules 17

Figure 1.7: Illustration of HLA pockets and binding peptide residues in HLA-B*27 18

Figure 1.8: Structure of HLA nomenclature 19

Figure 1.9: Comparison of KIR receptor structures, highlighting their domain organization, length of cytoplasmic tail and presence of ITIM sequences 23

Figure 1.10. KIR haplotypes and their gene content variability 23

Figure 2.1: The data curation pipeline, the types of data that were extracted from each publication and the submission workflow developed within KDDB 37

Figure 2.2: Entity-relationship diagram of KDDB schema 38

Figure 2.3: Screenshots of the data submission pipeline within KDDB 40

Figure 2.4: KDDB can be accessed through the AFND homepage (http://www.allelefrequencies.net/) using the menu "KIR" and the submenu "KIR and disease associations" 42

Figure 2.5: KDDB homepage providing links for querying the database and to submit new studies 43

Figure 2.6: The query interface within KDDB, showing the additional detail about a given association study retrieved by following the hyperlink 43

Figure 2.7: (A) Percentage of studies stored in KDDB classified by disease type. (B) Percentage of studies stored in KDDB classified by continent 45

Figure 2.8: Geographical distribution of KIR studies according to the disease type investigated 46

Figure 2.9: Geographical distribution of KIR studies investigating autoimmune and infectious diseases 47

Figure 2.10: Distribution of individual KIR gene associations with susceptibility or protection to disease types 50

Figure 2.11: Distribution of KIR genes associations with susceptibility or protection to disease types according to their function 51

Figure 2.12: Proportion of reported KIR associations associated with susceptibility to diseases belonging to autoimmune, cancer and pregnancy complications classifications of disease types 53

Figure 2.13: Clustered heatmap of KIR gene frequencies across populations in AFND 54

Figure 3.1: Polymorphic residues on HLA-A, -B and -C molecules..... 64

Figure 3.2: Single HLA-A antigen mismatch representing an acceptable mismatch from an HLA epitope perspective (Donor: A*32 / Patient: A*01)..... 65

Figure 3.3: Screenshot of the HLA Epitope Registry database, showing a subset of class I epitopes and HLA alleles where they are present, subdivided in a category containing only alleles present in Luminex® SAB and another category including alleles absent in SAB assays..... 66

Figure 3.4: Screenshot of EpVix [125] option for epitope analysis of alloreactive antibody SAB profiles generating all possible reactive epitopes, with additional options to manipulate the MFI cut-off value and interactively highlight epitopes shared by different alleles 67

Figure 3.5: Comparison of the variables used in the application of HWP for biallelic loci or HLA epitopes 70

Figure 3.6: Example of all the steps involved in the calculation of population frequency of 65GK and 66K epitopes in England North West population, using both types of entry data (HLA raw data and allele/haplotype frequencies) 71

Figure 3.7: Entity-relationship diagram of EpFreq-DB schema 73

Figure 3.8: Algorithm to determine minimum HLA epitope set capable of explaining reactive alleles in Luminex® SAB assays 76

Figure 3.9: Correlation between HLA epitope frequencies calculated from HLA raw data and HLA haplotype frequency data for populations with sample size greater than 100 78

Figure 3.10: EpFreq-DB query page 80

Figure 3.11: 'Heatmap' view of EpFreq-DB query results 81

Figure 3.12: Heatmap of HLA-A and -B epitope frequencies in world populations 82

Figure 3.13: Correlation between epitope matching percentages in populations obtained using 'minimum' and 'full' epitope sets as input 87

Figure 3.14: Dotchart of populations showing any difference in matching likelihoods between minimum and full epitope sets 88

Figure 3.15: Correlations between population matching percentages obtained using allele and epitope sets as input 89

Figure 3.16: Point plot of epitope matching frequencies minus allele matching frequencies in populations, categorized by patients / MFI cut-off combinations 90

Figure 3.17: Dot chart showing only populations where Luminex® allele matching frequencies are higher than epitope matching frequencies 91

Figure 4.1: Nevirapine 2D (A) and 3D (B) chemical 95

Figure 4.2: A model for the interaction of HIV-1 RT with a representative NNRTI (UC781), to exemplify the functional interaction of HIV-1 RT and nevirapine 96

Figure 4.3: Comparison of C*04:01 frequencies observed in cases and controls from the Malawian cohort with frequencies from worldwide healthy populations from AFND ('World', 'Sub-Saharan', 'Black') 104

Figure 4.4: Docking of nevirapine to HLA-C*04:01 peptide binding region, where a total of 20 conformation modes were produced 106

Figure 4.5: Docking of 12-hydroxy-nevirapine to HLA-C*04:01 peptide binding region, where a total of 20 conformation modes were produced 107

Figure 4.6: HLA-C*04:01 residues contacting all nevirapine (A) and 12-hydroxy-nevirapine (B) modes predicted by docking 109

Figure 4.7: Nevirapine (A) and 12-hydroxy-nevirapine (B) first pose docked to HLA-C*04:01, highlighting the contacting residues 111

Figure 4.8: List of all contacting residues of predicted conformations of nevirapine and 12-hydroxy-nevirapine in C*04:01, separately for each ligand cluster referring to a putative binding site 113

Figure 4.9: World distribution of allele frequencies of HLA-C alleles present in the Malawi cohort 114

Figure 4.10: *In silico* nevirapine binding to the crystallographic structure of the interaction between HLA-C*04:01 and KIR2DL1, highlighting the residues in both molecules contacting nevirapine 117

List of Tables

Table 1.1: KIR molecules and their respective ligands, signalling and function 25

Table 1.2: KIR-ligands in HLA- and HLA-B. HLA-C alleles are of the C1 or C2 group while most HLA-B alleles are Bw4 or Bw6, according to their motifs 26

Table 2.1: Comparison of KIR gene associations with susceptibility and protection to diseases reported in multiple association studies found in the literature, grouped by disease type 49

Table 3.1: Populations having HLA raw genotyping data used to generate epitope profiles 74

Table 3.2: HLA epitope analysis of three alloreactivity profiles from EpVix, using 2000 and 4000 MFI cut-off values 84

Table 4.1: HLA polymorphisms selected for nevirapine docking, including the C*04:01 and other HLA control molecules 102

Table 4.2: Polymorphic residue difference between C*04:01 and chosen control alleles within the peptide binding region protein sequence 102

Table 4.3: Comparison of HLA-C allele frequencies between controls and SJS/TEN patients in the Malawi cohort 103

Table 4.4: Polymorphic residues among the alleles present in Malawi cohort for which more than 50% of the alleles are different than C*04:01 and their HLA carrier frequencies 110

Chapter 1

Investigations in Immunogenetics

1.1 The human immune system

The human immune system is a complex system composed of cellular and humoral (extracellular macromolecules) components whose primary function is to defend the organism against pathogens. Its components and mechanisms are divided in two interrelated subsystems. The *innate immune system* is responsible for an initial and immediate response during an encounter with pathogens being able to recognize foreign molecules generally present in pathogens and to detect abnormal cells infected by viruses or transformed in tumour cells. The *adaptive immune system* launches a posterior, but more effective and specific response, which is also capable of "remembering" an infection, being able to quickly release the more effective response in case of re-infection. Although the primary function of the immune system is fighting foreign pathogens, it also interacts with other physiological mechanisms, such as pregnancy, and can be influenced by various other factors such as sleep and diet.

A characteristic of the human immune system is the remarkable genetic variability, innately present in germline genes or acquired by somatic recombination, of key molecules participating in the detection of invading microorganisms. This high molecular diversity is essential for the capability of the immune system to react and adapt its response to a wide range of ever evolving pathogens. The high number of immunogenetic polymorphisms maintained in the human population results in significant differences in the genetic makeup of immune genes between individuals, generating a consequent inter-individual variability in the manifestation of the immune system response against various microorganisms or tumours. Some of this variation can also lead to malfunctions of the immune response and cause autoimmune diseases, where the mechanisms that prevent the organism to react against itself get disrupted. Inter-individual difference of immune genes is also a crucial factor to be assessed prior to transplantation procedures to avoid elicitation of immune responses caused by antigens expressed by immune genes which are different from the ones possessed by the patient.

Although the immune system is conventionally subdivided in innate and adaptive immune systems, all the cells participating in the immune system are leukocytes originated from progenitor stem cells in the bone marrow. After maturing into diverse cell types, they migrate to peripheral tissues or circulate in the blood stream and lymphatic system to exert functions related to innate, adaptive or both systems. Leukocytes such as macrophages, basophils, neutrophils, eosinophils, mast cells and natural killer cell lymphocytes are a part of the innate immune system, while antigen specific B and T lymphocytes are part of the adaptive immune system. Another leukocyte, the dendritic cell, is considered to be a cell that interfaces the innate and adaptive immune systems, since its main function is to activate T cells by presenting them pathogen derived peptides and providing signals necessary for activation (Figure 1.1) [1].

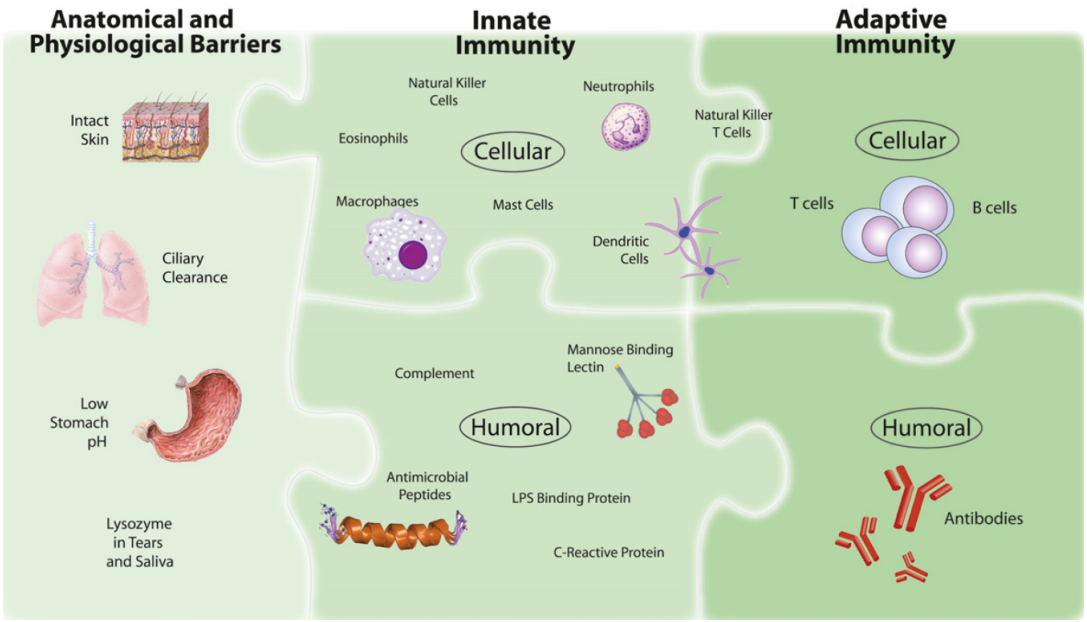


Figure 1.1: Components of the human immune system and the integration of innate and adaptive systems, from [2].

1.1.1 Innate Immune System

The innate immune system, also known as non-specific immune system, is the first line of defence against pathogens. It is composed of physical barriers, such as the epithelial surface, and of leukocytes and humoral molecules capable of recognizing generic patterns present in microorganisms. Regarding its cellular components,

phagocytes, such as macrophages, neutrophils, and dendritic cells, detect the presence of pathogens using pattern recognition receptors (PRRs), which are receptors that can recognize conserved molecular sequences commonly shared by pathogens. Those receptors are expressed by fixed genes innately encoded in the genome, differing from B and T cell receptors of the adaptive immune system, whose genes undergo multiple somatic genetic rearrangements to achieve an extremely high diversity, and therefore recognize and adapt a specific immune response [2]. PRR activation leads to the release of inflammatory mediators, resulting in the clinical symptoms of inflammation such as increased sensitivity of pain receptors and vasodilation in the inflammation site (redness). Those mediators also attract more immune cells to the inflammation site, especially phagocytes, which is also facilitated by the increased blood flow [1].

Phagocytes engulf and digest microorganisms into smaller amino acid sequences called peptides, which are then presented as antigens to cells of the adaptive immune system. Another name for phagocytes is antigen-presenting cells (APC), since this function is a link for the activation of the adaptive immune response. Through this mechanism, APCs educate lymphocytes from both the adaptive and innate immune system about specific infections, mainly from extracellular origin [3]. Although various leukocytes can perform this function, dendritic cells are specialized APCs. Their distinct morphological and functional characteristics gives them unique roles essential for stimulating, regulating and enhancing other cells of the immune system [4].

While the phagocytes are responsible for neutralizing microorganisms outside cells and warn the adaptive immune system, another group of leukocytes in the innate immune system, natural killer (NK) cells, are responsible for eliminating cells that have been compromised by intracellular pathogens, and are also able to eliminate tumour cells. Humoral defences include the lipopolysaccharide (LPS) binding protein, which plays a role in the mechanism for PRRs recognition, and the molecules composing the complement system whose function is, through various pathways, releasing a biochemical cascade resulting in the destruction of the microorganism [2].

1.1.2 Adaptive Immune System

The components of the adaptive immune system function to deliver a highly specific immune response against pathogens, further developing an immunological memory,

allowing its quick activation in case of a reinfection. Activation of this system is dependent of antigen exposure, which is one of the characteristics distinguishing it from the innate immune system. Despite its high efficacy, the adaptive immune system has a delayed response, taking up to 5 days for proliferation and differentiation into effector cells able to deliver a response [2]. In evolutionary terms, the adaptive immune system is more recent than the innate immune system, since it is only present in vertebrates [5].

T lymphocytes or T cells are mainly responsible for cellular immune responses while B lymphocytes or B cells are responsible for humoral immune response through the production of antigen-specific antibodies, also called immunoglobulins. Different from cell receptors of the innate immune system encoded from germline genes, T and B cell receptors (TCRs and BCRs) undergo a type of somatic gene recombination during their maturation in the bone marrow, which is referred as V(D)J recombination. Receptors are built by random assembly of V, D and J gene fragments, resulting in a hugely diverse receptor repertoire that enables those cells to recognize antigens from the vast amount of possible invading pathogens [6]. After acquiring their receptors, lymphocytes capable of reacting against self-molecules are eliminated, a maturation stage that occurs in the thymus for T cells and in the bone marrow for B cells. Immunocompetent cells then migrate to lymphoid tissues as naïve cells, ready to be exposed to antigens, causing it to proliferate by clonal expansion and differentiate into effector T cells or antibody-producing plasma B cells. TCRs interact with antigens in the form of peptides presented by cells through molecules of the major histocompatibility complex (MHC), while BCRs interact with their cognate antigens outside cells [1].

T cells are subdivided into helper and cytotoxic T cells, according to their different functions and the presence of distinguished TCR co-receptors (CD4 and CD8, respectively). These co-receptors determine the type of MHC molecule that the TCR can bind. Helper T cells or CD4+ T cells interact with MHC class II molecules present in APCs while cytotoxic T cells or CD8+ T cells interact with MHC class I molecules present in almost all cells [7]. Those differences are related to the fact that CD4+ T cells are mainly involved in the stimulation of immune mechanisms against microorganisms that have undergone phagocytosis by APCs while the function of CD8+ T cells is to eliminate cells that have been infected by pathogens. While cytotoxic T cells destroy infected cells presenting foreign peptides, helper T cells are responsible for releasing molecular signals, called cytokines, which mediate the activity of other cells, including cells of the innate immune system.

The activation of B cells via interaction of their BCRs to a cognate antigen initiates the humoral portion of the adaptive immune system. When activated, they differentiate to plasma B cells whose function is to secret immunoglobulins which are structurally similar to the BCR of the activated B cell. Immunoglobulins bind directly to invading microorganisms, coating their surfaces in order to neutralize them, i.e. prevent them from infecting cells. This antibody coating, referred as opsonisation, also facilitates their capture by phagocytes and is involved in pathways for activation of proteins of the complement system [1].

1.1.3 Immunoglobulins: Humoral effectors of the adaptive immune system

TCRs, immunoglobulins and BCRs are members of the immunoglobulin superfamily (IgSF) sharing mechanisms to increase diversity, such as V(D)J recombination, and structural similarities, being composed of a variable region (V) related to antigen recognition, and a constant region (C) related to activating of immune responses. Distinctively, immunoglobulins display an additional mechanism to increase their diversity. After gene rearrangement they undergo somatic hypermutation, consisting of point mutations in V regions that contribute to the increasing affinity of immunoglobulins during the course of the infection [1]. All these mechanisms make immunoglobulins highly specific to their targets. The understanding of the mechanisms governing the immunoglobulin specificity was essential to the development of structural based strategies to improve tissue transplantation, discussed in Chapter 3 of this thesis.

Immunoglobulins are Y-shaped molecules composed of two heavy (H) chains and two light (L) chains connected by disulphide bonds, where both H and L chains are equal. They are identical to BCRs from their originating cells, except for a small portion of the COOH-terminus of the H chain connecting it to the B cell membrane. The V region comprises of the N-terminal variable domains of the L and H chains (V_L and V_H, respectively), while the C region comprises of L and H chain conserved domains (C_L and C_H, respectively (Figure 1.2) [1].

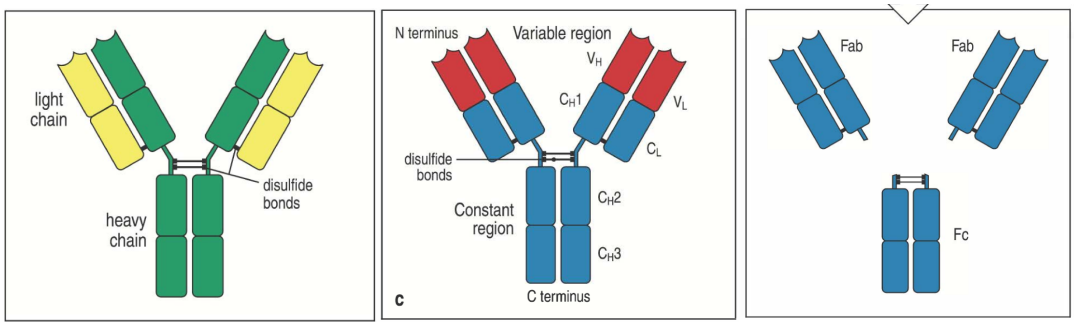


Figure 1.2: Immunoglobulin structure from different perspectives. (A) Highlight of immunoglobulin L and H chains. (B) Description of C and V immunoglobulin sub regions, according to L and H chains. (C) Indication of the Fab and Fc portions generated by proteolytic cleavage (Adapted from [1]).

V domains of both L and H chains contain three regions with a distinctive higher variability called hypervariable regions (HV) - HV1, HV2 and HV3, where HV3 is the most variable region. Those HV regions are located close to each other in the antigen binding site of the immunoglobulin, and are commonly referred as complementarity-determining regions (CDRs), as their surface structure matches the antigen surface [1]. Thus, both chains contain three CDRs (CDR-L1, -L2, -L3 pairing with CDR-H1, -H2 and -H3, respectively) where the total of six CDRs are brought together in each antigen binding site of the immunoglobulin, conferring its specificity (Figure 1.3) [8].

Immunoglobulin CDRs are encoded by recombinant V, D and J genes in distinct loci specific to L and H chains. While both CDR1 and CDR2 are singularly encoded by V gene segments in L and H loci, CDR3 displays increased variability by being encoded by combinations of V, D and J segments (CDR-L3 from V and J segments and CDR-H3 from V, D and J segments). Variability is especially increased in CDR-H3, since H chain locus have the additional polygenic D segment, greatly increasing the number of possible specificities generated in comparison to other CDRs. Furthermore, N-nucleotides (non-template encoded) flanking both sides of the D segment are randomly added by the TdT enzyme [8]. Those characteristics make CDR-H3 the most variable CDR and the dominant region on the determination of immunoglobulin specificity and affinity, while others CDRs do not variate as much and their structure is limited to a small set of main-chain conformations referred as canonical structures [9].

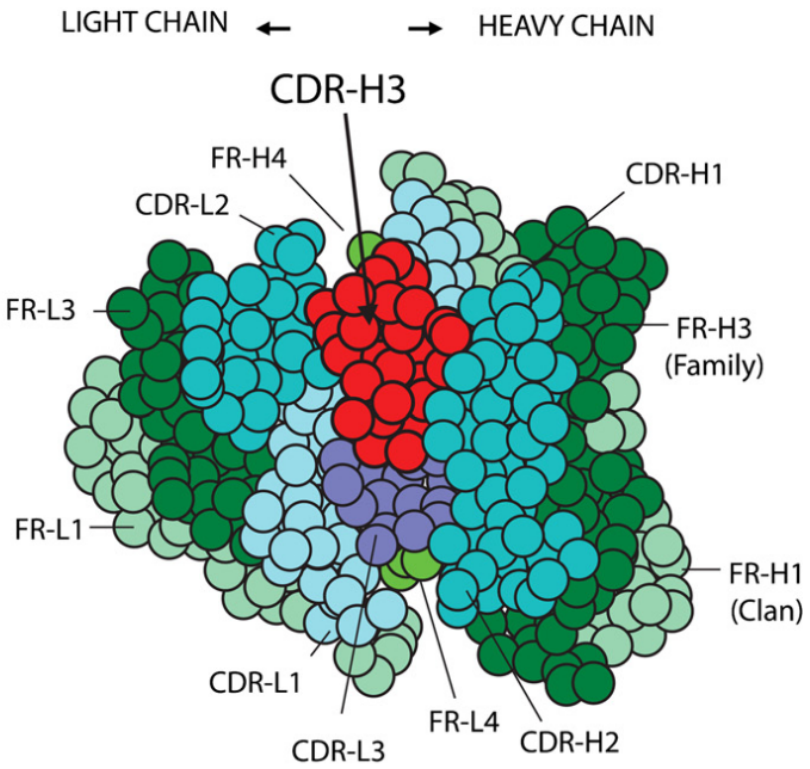


Figure 1.3: Complementary Determining Regions (CDR) loops in an antigen-binding site. CDR-H1, -H2 and -H3 are located in the antibody heavy chain, and L1, L2 and L3 are located in the antibody light chain. The centrally located CDR-H3 confers specificity to the antibody, while interactions with other CDRs are secondary (Adapted from [8]).

Nevertheless, it has been shown that among the 50 HV amino acid residues in the immunoglobulin V region, small numbers of residues make contact with the protein antigen, interfacing around 15-22 antigenic amino acid residues [10]. This antigen contact surface is the structural definition of an *epitope*. An antigenic molecule can carry several epitopes, consisting of small immunogenic sub regions displaying non-self protein conformations. The epitope complementary region in the immunoglobulin is called *paratope* [8]. An epitope within a protein can be formed by non-linear amino acids structurally together due to protein folding, originating conformational or discontinuous epitopes. The opposite case are continuous or linear epitopes, where the epitope is comprised of amino acids linear sequence in the polypeptide chain [1].

Knowledge of general epitope structural conformations (linear and non-linear) recognized by CDR-H3 allowed for prediction of HLA epitopes that could play a role in antibody-mediated transplant rejection. Areas over the surface of an HLA molecule containing polymorphic residues that are within predicted area recognized by CDR-H3 were defined as functional units important for matching in transplantation, providing new alternatives for current DNA-based matching. Definition of HLA epitopes is discussed in more detail in Chapter 3 of this thesis.

1.1.4 Natural Killer Cells: Lymphocytes of the innate immune system

Despite being categorized as part of the innate immune system since they do not require exposure to foreign antigens and their receptors do not undergo somatic recombination, NK cells are lymphocytes originated from progenitor cells common to T and B lymphocytes. Initially considered to be primitive lymphocytes, it has been suggested they have co-evolved alongside lymphocytes of the adaptive immune system as a response to pathogens that downregulate MHC class I molecules to evade from cytotoxic T cells [11-13]. Although NK cells were discovered as a result of their ability to target and kill tumour cell lines that expressed little or no HLA class I molecules [14]. It is now known that the killing function in NK cells is dependent on a mixture of activating and inhibitory receptors present on the membrane and the interaction with their respective MHC ligands [15].

Normal cells express a high quantity of MHC class I molecules on their surface while low or absent MHC class I expression are a sign of infection or neoplastic transformation. Equipped with a complex of receptors that monitor ligand expression in cells, NK cells are capable of identifying and destroying cells with abnormal levels of MHC class I expression, a scenario also referred as "missing-self". In summary, NK cell interaction with other cells, also referred as NK synapse has the following stages: i) interaction of receptors with respective ligands in target cell; ii) integration of signals from bound receptors; iii) NK cell remains inactivated or becomes activated; iv) if activated, NK cell destroys the target cell by releasing cytotoxic granules and v) release of cytokines that modulate activity of other cells. Several receptor families in NK cells recognize MHC class I ligands, and some have the function to send inhibitory signals that prevent NK

cell activation. In the presence of insufficient inhibitory signals, NK cells are activated by the interaction of activating receptors with activating ligands [16].

Inhibitory signals are released via phosphorylation of a tyrosine residue in the immunoreceptor tyrosine-based inhibitory motif (ITIM) cytoplasmic domain, present in all NK inhibitory receptors, which then leads to recruitment of phosphatases through their SH2-domains that act preventing NK cell activation. Signalling by activating receptors occurs via phosphorylation of a tyrosine residue in the immunoreceptor tyrosine-based activating motif (ITAM), present in NK cell adapter proteins associated with NK cell activating receptors. Tyrosine kinases are then recruited to initiate a cascade of events leading to Ca^2+ influx, degranulation, and transcription of cytokine production [11].

Receptors families in NK cells recognizing MHC class I or related ligands include C-type lectins-like group (CD94/NKG2), killer-cell immunoglobulin-like receptors (KIR) in humans and *Ly49* in mice and LILR. Despite most of the NK cell receptors binding MHC class I related molecules, several receptors bind non-HLA ligands, for example, CD16 binds IgG, triggering an activating response, and NKp30, NKp44 and NKp46 are activating receptors that bind molecules expressed by pathogens and self-ligands. Among those receptors, KIRs are the most variable family. They are expressed by a variety of KIR genes, resulting in several molecular variants of KIR receptors that, despite being similar, can perform different functions (inhibitory or activating), however most KIRs have inhibitory function [17-21].

Further discoveries after the initial description of NK cell physiology led to updates in the "missing-self" hypothesis (Figure 1.4). The initial hypothesis stated that NK cells are activated when encountering cells lacking MHC class I ligands. Nowadays it is known that NK cells can also be activated by cells overexpressing activating ligands with no MHC class I downregulation. Furthermore, cells that naturally lack MHC class I ligands, like erythrocytes, also lack activating ligands for NK cells and do not elicit an activating response on those cells [11]. Variability of both NK receptors and their respective ligands in target cells also represents a factor interfering in the determination of NK cell activation. Due to the extensive polymorphism on some NK cell receptor genes, individuals display variable receptor-ligand repertoires therefore having variable balances of potential inhibitory and activating signals. This variability can also produce a

deleterious effect in situations where an excess of activating signals due to a more "activating" combination of receptors results in autoimmune diseases [22].

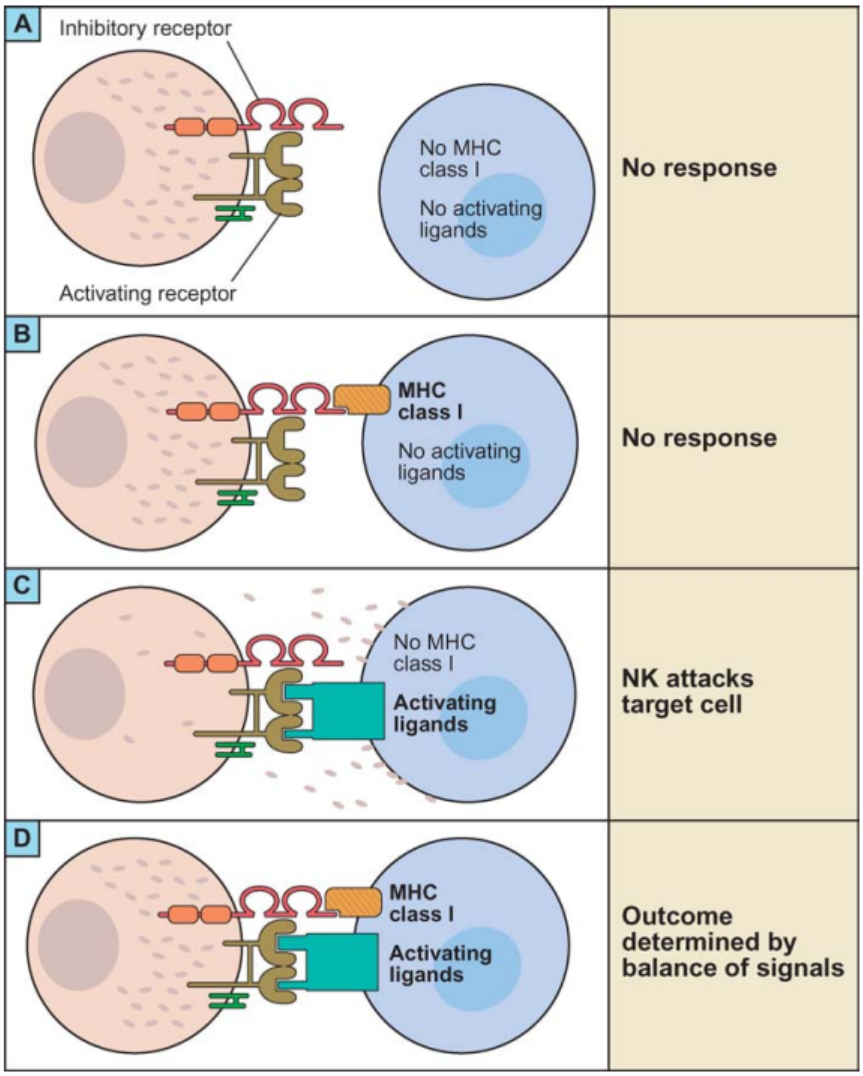


Figure 1.4: Mechanisms determining NK cell activation according to an updated version of the "missing-self". The classical definition of this hypothesis states that NK cells becomes activated when targeting cells missing MHC class I molecules. Updates on this hypothesis includes other possible scenarios. Erythrocytes are examples of cells which do not express MHC class I ligands, but do not lead to NK cells activation as it lacks activating receptors (A). Cells that in normal conditions express MHC class I lead to inhibitory signals on NK cells (B), but infection or tumour transformation can lead to downregulation of MHC class I and expression of activating ligands (C). Some cells normally express both MHC class I and activating ligands, and can overexpress activating ligands in unhealthy situations with no changes in MHC class I expression, shifting the balance of signals towards NK cell activation (D). Variability in individual NK cell receptors-ligands repertoire also influences the balance of signals and thus the sensitivity of NK cells [11].

NK cells also participate in mechanisms related to pregnancy. Uterine NK (uNK) cells are the most abundant leukocytes in the uterus, and regulate the invasion of the trophoblast into the uterine wall and the transformation of maternal arteries in high conductance vessels [23]. Similarly, different receptor combinations influence this mechanism, in this case mainly due to interaction with paternal ligands present in the trophoblast. Excessive inhibition of uNK cell receptors by trophoblast ligands have been suggested to impair transformation of maternal arteries, leading to pregnancy disruption [24].

Knowledge of the mechanism of NK cell activation, relying on its receptor repertoire, combined with target cell ligands can also be used for therapeutic strategies. Using allogeneic NK cells for cancer immunotherapy is one strategy that takes advantage of imbalances on NK cell receptor signals, where differences in KIR receptors and MHC ligands between recipients and donors cause NK cells to be more or less aggressive against tumour cells in the recipient [25]. In transplant settings, NK cells can influence the outcome both positively and negatively. For bone marrow transplants, they play an important role on preventing cancer relapse by targeting tumour cells, but they can also participate on immune responses against the host [26]. In solid organ transplantation, they can be involved on graft chronic rejection [27].

Since the polygenic and polymorphic nature of KIR receptors produces a high genotype diversity in populations compared to other NK cell receptors, they have been implicated in shaping individual response to infections and tumours, in the outcome of transplantation, and in the occurrence of autoimmune diseases and pregnancy complications. Furthermore, different KIR receptors display different levels of affinity to their ligands, which is also influenced by ligand polymorphisms, i.e. different MHC alleles [24]. The influence of human KIR variability in disease development is investigated in this thesis (Chapter 2).

1.2 Genetics of the immune system

Genetic polymorphism refers to nucleotide differences in one or more positions in the DNA sequence between individuals. The sole term *polymorphism* can also refer to their phenotypic variants, i.e. their expressed proteins. Sometimes gene polymorphisms are

referred as *alleles*, which are alternative genetic variants that can be present at a position (or *locus*, plural *loci*) in the chromosome. The allele definition is similar to genetic polymorphism, the difference being that the term allele is inferring chromosome *diploidy*, i.e. the fact the humans have chromosome pairs, each one inherited from one of the parents, and therefore a double set of each gene in a particular locus, termed *genotype*. Thus it is possible for individuals to express two different alleles of a polymorphic gene (*heterozygosity*).

Part of the ability of the human immune system to identify and react against a broad range of constantly evolving pathogens that can infect the organism is interrelated with the extensive diversity displayed by a set of immune gene families. The Human Leukocyte Antigen (HLA) and the Killer-cell Immunoglobulin-like Receptors (KIR) gene families present an outstandingly high variability and express key components related to monitoring for the presence of non-self peptides or abnormal ligand expression on the cell surface, respectively. Both exhibit elevated number of polymorphisms, HLA being situated in the most polymorphic locus in the human genomic, while KIR also varies in haplotype gene content. The consequent variability of immunogenetic profile in populations and the functional implications of the molecules expressed by those genes makes them frequent targets in disease association studies investigating the correlation of immune gene polymorphisms with the observation of different disease outcomes, i.e. a higher susceptibility or protection against different disease types.

A set of genetic polymorphisms inherited together on a region of a single chromosome, from a single parent, are called *haplotypes*. For a given chromosomal region, individuals have a set of two haplotypes (one on each chromosome), and the combination of the polymorphisms in both haplotypes is also referred as *genotype*. The HLA gene complex is located in the chromosome 6, while KIR gene complex is located in the chromosome 19. Therefore, genes in each of those gene families respectively are linked in the same chromosome, resulting in different haplotypes with variable frequencies in populations.

1.2.1 Population Genetics

Study of genetic polymorphism frequencies in populations is a component of *population genetics* research. The investigation of genetic frequencies is important for

understanding evolutionary biology, population dynamics, and also disease association studies. Several population genetics measures are commonly used in HLA and KIR population and disease association studies. A common one is *allele frequency* which consists of the percentage of an *allele* polymorphism in a population chromosome pool. Since individuals have two chromosomes, they can carry one or two copies of the allele. Therefore, the *allele frequency* reflects the proportion of the *allele* in all chromosomes in the population. This is different from another measure referred as *carrier frequency*, which is the percentage of individuals in the population carrying a polymorphism, disregarding zygosity (heterozygous or homozygous). *Carrier frequency* is commonly used in KIR gene studies, due to their variable gene content of KIR haplotypes (presence / absence) making it difficult to determine their zygosity.

Regarding genotype information, an individual can have two copies of the same allele (*homozygous*) or different alleles in each chromosome (*heterozygous*). The *genotype frequencies* for a locus consists of the percentage of genotypes from individuals in a population, homozygous and heterozygous. For a biallelic locus, i.e. a locus containing two polymorphisms (*A* and *a*), three genotype frequencies (*AA, Aa* and *aa*) can be estimated. *Genotype frequency* can also denote the frequencies of multi loci genotypes, where all genotypes in a population correspond to the pairwise combination of all existing haplotypes in that population. In this case, the equivalent of the allele frequency would be the *haplotype frequency*.

Hardy-Weinberg Equilibrium

Genotype frequencies can be estimated according to probabilistic proportions based on the population allele frequencies, which are referred as Hardy-Weinberg proportions. If the observed genotype frequencies in a population are statistically equivalent to the expected by Hardy-Weinberg proportions (HWP), the population is in Hardy-Weinberg equilibrium (HWE) [28]. Although HWE can be affected by natural selection and population events, shifts in HWE are often indicative of problems with population sampling or typing methods [29]. HWP is based on the probabilities of the alleles in a population to be homozygous or heterozygous. Using again a biallelic locus as example, *A* and *a* alleles have *p* and *q* frequencies, respectively, in a population. According to probabilistic theory, p^2 and q^2 are the expected frequency of homozygotes (*AA* and *aa*),

while *2pq* is the expected frequency of heterozygotes (*Aa*). This logic yields the general HWP formula:

$$p^2 + 2pq + q^2 = 1$$

HWP can also be applied for estimating other mathematical variables in populations. For example, a formula derived from HWP, sometimes referred as Bernstein's formula, is used to estimate gene frequencies from carrier frequencies, which is useful for KIR studies [29]. In this thesis, another calculation derived from HWP has been applied to estimate HLA epitopes frequencies, described in Chapter 3.

Linkage Disequilibrium

Haplotypes are also expected to be proportionally balanced or in linkage equilibrium. Linkage disequilibrium can be caused by factors such as recombination rate and genetic linkage, i.e. their simultaneous presence in the same chromosome [28]. Tightly linked loci are more likely to be in linkage disequilibrium, as they have a higher chance of being inherited together [30]. HLA genes have a very high genetic linkage, being clustered on a relatively small region of chromosome 6, and therefore display high linkage disequilibrium among themselves. The same is true for KIR genes, clustered on a on chromosome 19. If natural selection acts on a single locus in highly linked gene families, its frequency is increased together with other closely linked loci in the haplotype, even if they are neutral regarding selection [29,31]. This phenomenon represents a challenge for genetic studies to identify the locus (or loci) actually being selected or being associated with a particular condition. Studies associating genetic variants with traits or diseases tend to report all statistically significant associations found. Some of the genetic variants associated may not be directly influencing the investigated condition, but show statistical significance by being in linkage disequilibrium with an actual causative variant. Since both HLA and KIR gene families display high degree of genetic linkage, it is necessary to take linkage disequilibrium into account before drawing definite conclusions from genetic association studies of those gene families [32].

Haplotype Frequency Estimation

To assess the effects of linkage disequilibrium in population genetic studies it is necessary to acquire haplotype frequencies of the studied loci, but most genetic studies

rely on DNA typing techniques that only provides genotype information, which is not enough to determine what polymorphisms are linked in the same chromosome. This is because most techniques rely on generating information on relatively short sequences of DNA (either by direct sequence or PCR), up to a few kb in length. As a result, phasing between alleles from different loci cannot be easily achieved, due to the typical presence of identical sequences in-between on both chromosomes typed. One method to obtain haplotypes is to obtain genotype data from the individuals' parents, but this is often impractical. It is possible, however, to use statistical methods to estimate the haplotypes and their frequencies in a population based on population genotype data. There are several computational methods for achieving this task, including some software packages designed specifically for either HLA or KIR gene families. For the level of polymorphism exhibited by those gene families, maximum likelihood methods using the expectation maximization (EM) algorithm are mostly used [32]. These methods are reasonably effective for large populations, say > 1000 individuals, working on the likely co-occurrence of the same alleles in different individuals, but start to break down for low frequency haplotypes in smaller populations, as there is simply no signal with which to work. A module for haplotype and linkage disequilibrium estimation from HLA genotypes using this method is present in PyPop (www.pypop.org), which is a software package containing an array of tools designed for population genetics analysis of highly polymorphic multi loci genetic data [33].

1.3 The Major Histocompatibility Complex (MHC)

The major histocompatibility complex (MHC) is a genomic region of approximately 4 Mb, located at the 6p21.3 region of the human chromosome 6. It contains more than 200 closely linked genes, performing various functions, mostly related to the immune system [34]. It is by far the most polymorphic region in the human genome, where many of its genes exhibit thousands of known allelic variants. As shown in Figure 1.5, the complex is divided into three regions (class I, class II and class III). MHC class I and class II regions are mostly composed of HLA genes while the class III region, located between class I and II regions, is composed of non-HLA genes with immune-related functions. The twenty one highly polymorphic genes of the HLA gene complex have been constant subjects of interest in diverse fields of study. Besides influencing disease susceptibility and immune reactions such as adverse drug reactions, their variability also makes them

important assets in genetic studies investigating human population origins and dynamics, and a problem in organ and cell transplantation as differences in donor and patient HLA profiles can lead to immune mediated tissue rejection.

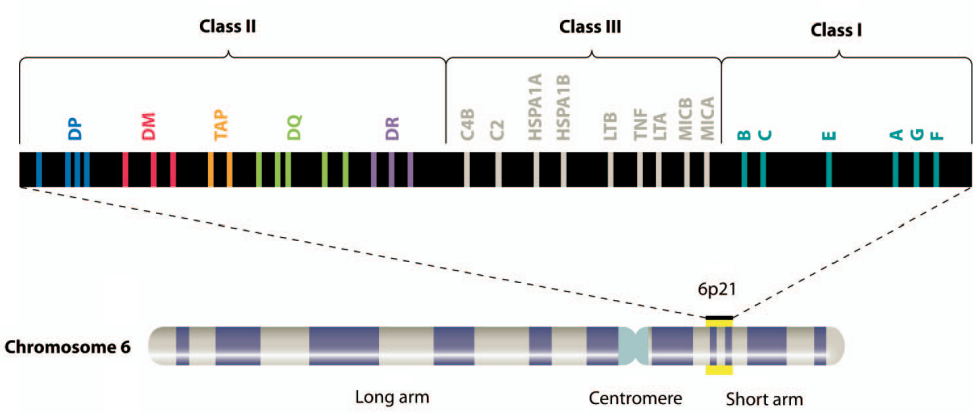


Figure 1.5: Location of MHC in chromosome 6 and genetic organization of class I, class II and class III regions. HLA genes are located in class I (HLA-A, -B, -C, -E, -F and -G) and class II (HLA-DR, -DQ, -DP and -DM) regions [Adapted from [35]].

1.3.1 Human Leucocyte Antigens (HLA)

The HLA gene complex is located within the MHC region. HLA class I (HLA-A, -B, -C, -E, -F and -G) and class II (HLA-DM, -DO, -DP, -DQ and -DR) genes are codominant, and translate for cell surface molecules responsible for antigen presentation to T cells. HLA class I genes are expressed in almost all cells, and their main function is to present peptides from intracellular microorganisms to cytotoxic (CD8+) T cells. HLA class II genes are mainly expressed in APCs, such as macrophages and dendritic cells, and their main function is to present peptides that have been phagocytized by APCs to helper (CD4+) T cells [35,36].

HLA class I molecules are composed of a heavy α-chain encoded by HLA class I genes, a light chain called β_2-microglobulin (β_2M) and a short bound peptide (Figure 1.6). Human β_2M is an extracellular molecule encoded by a conserved gene on the chromosome 15. The α-chain is composed by different domains encoded by specific exons of HLA class I genes. The leader peptide is encoded by exon 1, the α1, α2 and α3 extracellular domains are encoded by exon 2, 3 and 4, respectively, the transmembrane portion is encoded by exon 5, the cytoplasmic tail is encoded by exons 6 and 7, and exon

8 is the 3' untranslated region (UTR). Despite sharing domain structure similarity with HLA class I and having analogous function, HLA class II molecules are composed of two transmembrane chains (α chain and a β chain), both encoded by HLA class II genes, and a bound peptide (Figure 1.6). Similarly to HLA class I molecules, each exon encodes for specific portions of each chain, but HLA class II genes encode for only two extracellular domains (α1 and α2 for α chains, β1 and β2 for β chains) from exons 2 and 3, respectively. Thus, each HLA class II molecules is composed of four extracellular domains [36].

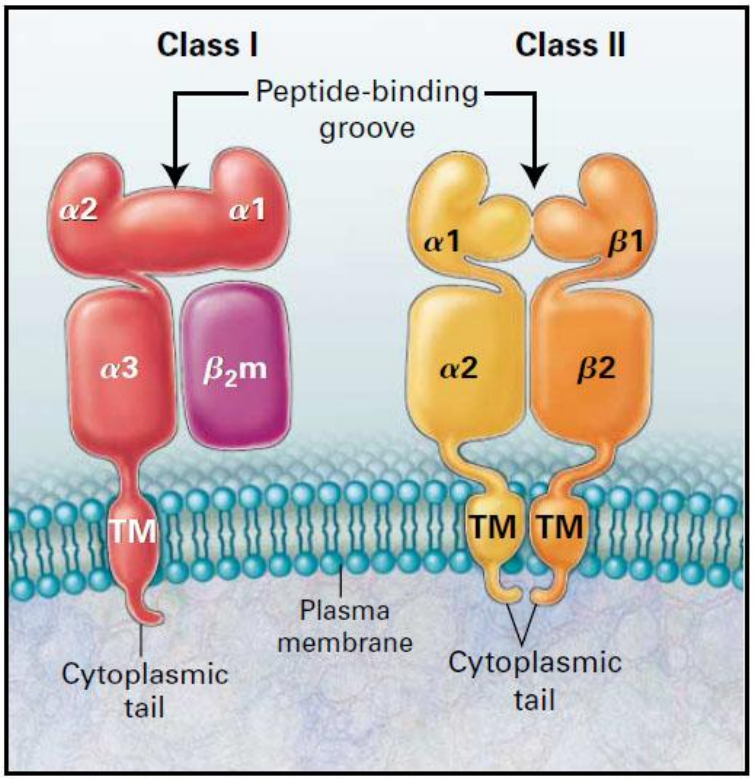


Figure 1.6: Structure of the HLA Class I and Class II molecules [37].

Membrane distal domains of HLA class I and class II molecules (α1 and α2 for class I, α1 and β1 for class II) form a groove where the peptide is tightly bound. This peptide-binding groove (PBG) consists of a sheet of N-terminal β-pleated strands forming the "floor" of the groove, and COOH-terminal α-helices forming its "walls". The top of the molecule comprising the peptide and α-helices is the surface that interacts with TCRs and other receptors. HLA class I peptides contain generally 8 to 10 amino acids, while HLA class II accommodates larger peptides usually from 12 to 24 amino acids as its PBG has

open ends, allowing peptides to hang out of the PBG. HLA class I PBG fully surrounds the peptide, limiting its size, and each extremity of the peptide is pinned to opposite ends of the PBG, mainly by hydrogen bonds. Every HLA molecule expressed in the molecular surface has a bound peptide originated from cellular degradation products. In healthy cells, the peptide source is from normal cellular components (self-peptides). In infected cells, some of the peptides are derived from pathogens [36].

Although thousands of different peptides can be bound to HLA molecules, they are not randomly selected, as subsites or pockets (Figure 1.7) of the PBG display preference for certain residues in specific positions of the peptide sequence, called peptide-binding motifs. Most polymorphic residues of HLA class I molecules are located within those pockets, influencing their size, shape and charge. Therefore, different HLA types have variable peptide sequence preference. While HLA class I pockets, mainly located in PBG extremities, primarily influence peptide affinity, HLA class II peptide affinity is determined by conserved residues centrally located in the PBG, resulting in lower peptide selectivity by HLA class II molecules. Therefore, polymorphisms within HLA class II PBG are mainly located in complementary pockets secondary to peptide affinity.

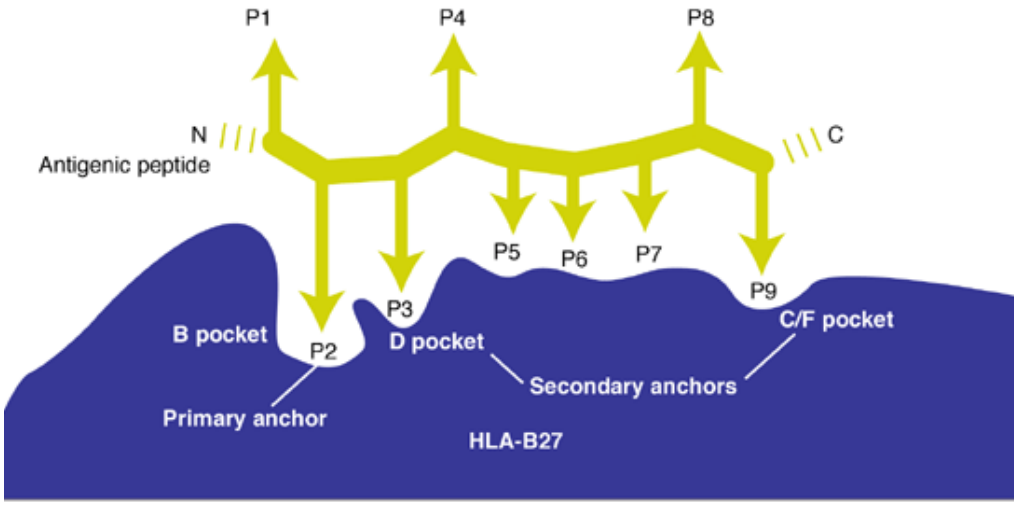


Figure 1.7: Illustration of HLA pockets and binding peptide residues in HLA-B*27 [38].

In HLA class I molecules most of the polymorphic differences in DNA sequence are in exons 2 and 3, coding for the most external domains (α1 and α2) located on the top of molecule. Hundreds of different alleles are known for HLA-A, -B and -C, while other HLA class I genes are monomorphic or display little polymorphism. Polymorphic

differences in HLA class II molecules can derive from both chains, depending on the isoform. HLA-DR α-chain (HLA-DRA) is monomorphic, while its β-chain (HLA-DRB) is polymorphic, being the most variable HLA class II isoform. HLA-DRB is a polygenic segment with variable gene content (HLA-DRB1, -DRB3, -DRB4 and -DRB5), where HLA-DRB1 is a highly polymorphic gene present in all haplotypes while other HLA-DRB genes can be present or not. In contrast, both chains of HLA-DP and HLA-DQ are polymorphic, while HLA-DM and HLA-DO show little polymorphism [36].

HLA nomenclature and resolution levels

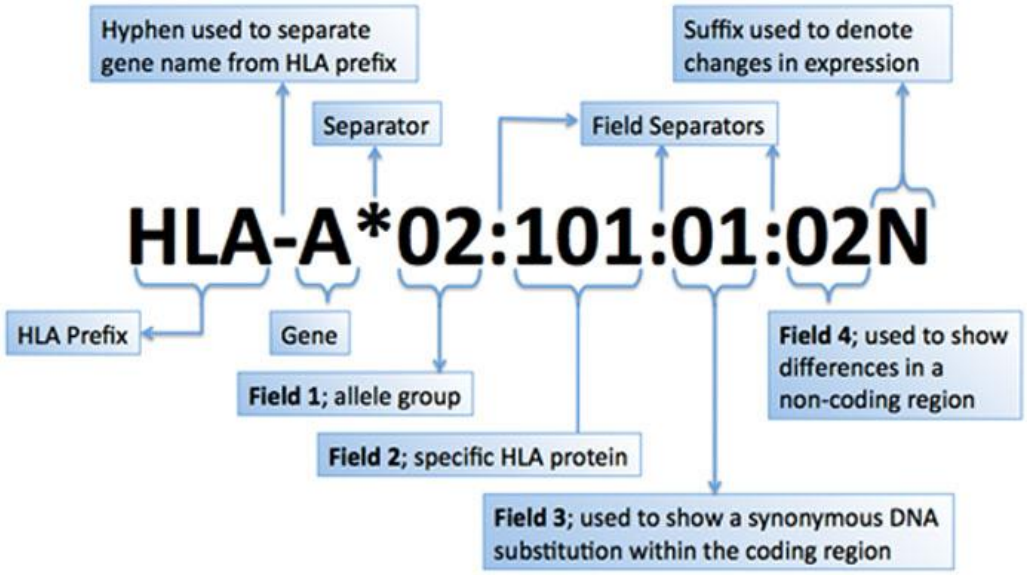


Figure 1.8: Structure of HLA nomenclature.

(Source: http://hla.alleles.org/nomenclature/naming.html)

HLA genes can be typed at different levels of resolution depending on the typing technique employed. Antibody-based serological typing yields low resolution allele definitions while DNA-based typing methods generates allele typing at higher resolutions. The resolution by which an allele has been typed can be identified by its nomenclature (Figure 1.8). The HLA nomenclature also provides information regarding protein expression. HLA allele names start with a prefix referring to the gene name. An asterisk separates the prefix from the rest of the nomenclature, which consists of up to four sets

of numbers called fields and a suffix that is present in some allele names. Fields are separated by colons, and represent different HLA resolution levels. The first field refers to the allele serological antigen designation. The second field corresponds to alleles differing in the amino acid sequence, but part of the same serological group. The third field classifies alleles with differences in the nucleotide sequence, but no difference in the protein sequence (synonymous substitutions). The last field is used to distinguish alleles that only display sequence differences in introns or untranslated regions. Some allele names can also contain a suffix, consisting of a single letter ('N', 'L', 'S', 'C', 'A' or 'Q') that refers to peculiarities in the allele expression. The suffix '**L**' indicates '**L**ow' cell surface expression in comparison to normal levels of allele expression. The '**S**' suffix indicates a soluble '**S**ecreted' molecule not present on the cell surface. The '**C**' suffix designates alleles expressing '**C**ytoplasmic' proteins not present on the cell surface. The '**A**' suffix indicates an '**A**berrant' expression, in cases where it is not known if the protein is expressed. The '**Q**' suffix refers to '**Q**uestionable', in cases where there is evidence that the mutation observed in the allele affects protein expression [39].

High resolution HLA typing methods can sometimes generate ambiguous HLA types. DNA sequence-based methods are often restricted to exons 2 and 3 (for HLA Class I genes), or exon 2 only (for HLA class II), since those exons encode for major determinants of the peptide-binding region specificity, containing most of the polymorphic positions. Differences between two alleles located outside the genotyped exons generates allele ambiguity. Another type of ambiguity referred as genotypic ambiguity occurs when it is not possible to distinguish between different heterozygous allele combinations without haplotype information. Since unambiguous HLA typing is limited to specialized laboratories, statistical methods using HLA frequency data from world populations are used for solving ambiguous HLA data [40].

HLA evolution and diversity

It is believed that the remarkable diversity displayed by HLA is an evolutionary response to various infectious diseases which have exerted a balancing selective pressure, favouring diversity in those loci and heterozygote excess in populations. This hypothesis is supported by the fact that most polymorphisms are observed within the 'peptide-binding site' (or antigen recognition site) region. Where polymorphism exists (in individuals or populations), a wider range of antigens can be recognized, suggesting also

that individuals heterozygote for HLA loci are more adapted to survive in pathogen-rich environments [41]. Due to the high number of HLA polymorphisms, most individuals are heterozygous for HLA genes, thus expressing six kinds of HLA class I molecules plus six or more kinds of class II molecules. Diversity of HLA genes is associated with antigen presenting functions, increasing the range of possible peptides to be presented, and also shaping the interaction strength between MHC and T cells. Thus, heterozygous individuals have more options of PBG shapes, increasing the range of possible peptides to be presented. Furthermore, other immune-related molecules also contribute to shaping HLA diversity. HLA class I molecules are known to interact KIR receptors, which are also very diverse. Although KIR genes are not linked genetically to the HLA complex, being on a different chromosome, it is suggested that both may have co-evolved due to certain combinations of HLA-KIR molecules causing complications in pregnancy [42,43].

The rapid evolution of human HLA genes generates extensive allele differences across populations. Those population specific alleles are often derived from more widely distributed alleles, differing only by small sequence segments that can be found in other alleles [44]. Besides making HLA genes a material for the study of the history of populations, this characteristic influences all other HLA related analysis due to how those genes can be genotyped. HLA typing in lower resolution levels do not cover the whole HLA gene sequence and may omit those population specific differences. Therefore, results from HLA association studies using low resolution typing methods may not be replicated in other populations if the causative molecular region in only distinguished by high resolution typing.

HLA association with human pathologies

Because of their immune function and their high level of polymorphism, HLA genes are obvious targets for disease association studies. They have been shown to influence individual susceptibility to several pathologies, mostly infectious and autoimmune diseases [35,45]. They have also been linked to complex diseases [46-48], neurological disorders [49-51] and adverse reaction to drugs [52]. Since the expressed HLA protein plays a role in various pathways of the immune response, different immunological mechanisms can be the underlying cause of those associations. HLA variants associated with those pathologies usually vary across different populations due not only to their diverse HLA allele pool and allele frequencies, but also due to variable environmental

pressures exerted in different populations [41]. It is also important to note that HLA may be associated with a pathology in conjunction with other genes. A classic case is the epistatic interaction between HLA class I and KIR genes, and several combinations of their variants have been linked to various disease types.

1.4 Killer-cell Immunoglobulin-like Receptors (KIR)

The KIR gene cluster is located in the leukocyte receptor complex (LRC) at position 19q13.4 of the human chromosome 19 [17,18]. To date, sixteen *KIR* genes have been identified, coding for receptors with activating (*KIR2DS1*, *KIR2DS2*, *KIR2DS3*, *KIR2DS4*, *KIR2DS5A/B* and *KIR3DS1*) or inhibitory (*KIR2DL1*, *KIR2DL2*, *KIR2DL3*, *KIR2DL5A*, *KIR2DL5B*, *KIR3DL1*, *KIR3DL2* and *KIR3DL3*) function, with *KIR2DL4* appearing to have both functions. Two pseudogenes *KIR2DP1* and *KIR3DP1* have also been identified [53]. KIR receptors are transmembrane glycoproteins members part of the IgSF, constituted of two (KIR2D) or three (KIR3D) Ig-like domains, a transmembrane region, and a cytoplasmic tail. Structurally, the activating and inhibitory functions of KIR are related to the length of their cytoplasmic tail that can be short (S) or long (L), distinguished in their nomenclature [53]. Long cytoplasmic tails contain one or two ITIM sequences, and short cytoplasmic tails have no ITIM sequence. KIR receptors with short cytoplasmic tails have a lysine in the transmembrane region, which is important for the receptor to interact with the DAP12 adapter protein that plays a role in NK cell activation [19].

Despite having distinct functions and variable ligand affinity, KIR receptors are structurally similar (Figure 1.9). They have a rapid evolutionary rate due to their close positions in the genome and their high sequence similarity, a consequence from gene duplications and conversions [54]. One example is the conversion of the inhibitory KIR2DL2 into the activating KIR2DS2, where two ITIMs are present in KIR2DL2, while in KIR2DS2 a stop codon in its cytoplasmic domain precedes the ITIMs, which are non-translated [11]. Variation in KIR loci can result from different gene and/or allele content of an individual [55], giving rise to haplotype diversity and leading to a very large number of different genotypes that have been observed (presence/absence of KIR genes). Its polygenic nature and haplotype gene content variation represents a challenge for genotyping KIR genes while discriminating zygosity information. Most KIR genotyping tests performed by laboratories are only able to reveal presence or absence of

KIR genes in populations, from where carrier frequencies can be calculated by direct count and gene frequencies need to be estimated using Bernstein's formula [29].

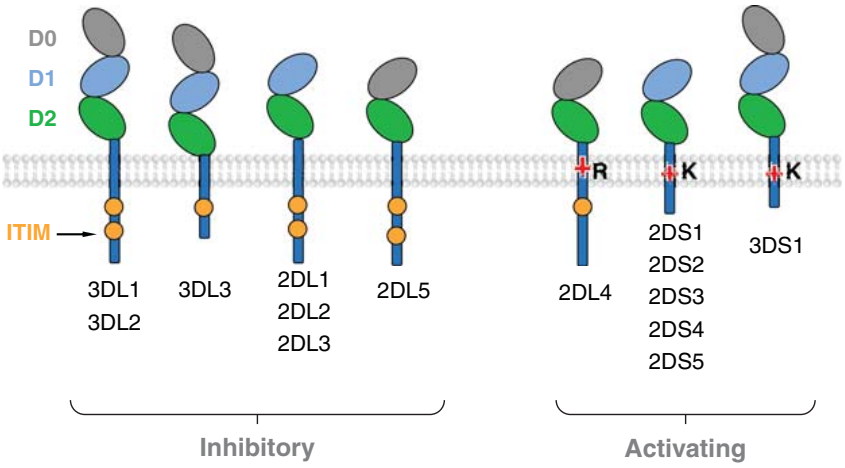


Figure 1.9: Comparison of KIR receptor structures, highlighting their domain organization, length of cytoplasmic tail and presence of ITIM sequences [19].

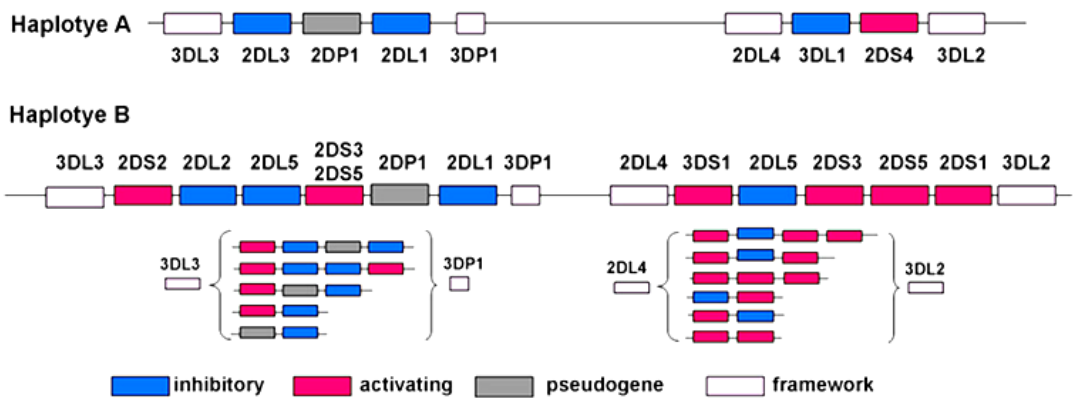


Figure 1.10. KIR haplotypes and their gene content variability. Haplotype A is generally non-variable in its gene content, while the B haplotype contains one or more of the genes encoding activating KIRs (KIR2DS1/2/3/5 and KIR3DS1) and the genes encoding inhibitory KIRs (KIR2DL5A/B and KIR2DL2) [Adapted from [24]].

The KIR genes *KIR2DL4*, *KIR3DL2*, *KIR3DL3* and *KIR3DP1* are present in nearly all individuals with a few exceptions [56], and are commonly known as 'framework' genes. The frequencies of inhibitory and activating genes vary in different populations, as

reviewed in [56]. A 24-kilobase band using *Hind*III digestion and Southern blot analysis distinguishes the haplotypes, termed A and B that make up the genotype [57].

The A haplotype is generally non-variable in its gene content – framework genes plus *KIR2DL1*, *KIR2DL3*, *KIR2DS4* and *KIR3DL1*, although occasionally one of these genes may be missing [56]. In contrast, the B haplotype contains one or more of the genes encoding activating KIRs (KIR2DS1/2/3/5 and KIR3DS1) and the genes encoding inhibitory KIRs (KIR2DL5A/B and KIR2DL2) (Figure 1.10). In B haplotypes variability is created by both the presence/absence of a gene and by allelic variation, in contrast, A haplotypes owe much of their variability to allelic variation [56].

According to the Allele Frequencies Net Database (AFND), a public database storing population variability of immune genes developed in our research group, more than approximately 600 KIR genotypes have been identified in worldwide populations [58]. Regarding allelic variability, 907 KIR alleles have been reported according to the last release of IPD-KIR (Release 2.7.0) [59], a public database storing KIR gene nomenclature and sequences following standards of the KIR Nomenclature Committee [60]. B haplotypes tend to be more prevalent in non-Caucasian populations, such as Australian Aborigines and Asian Indians, whereas in Caucasian populations approximately 55% will have one and 30% two A haplotypes [61,62]. It is thought that populations with higher frequencies of B haplotypes are those under strong pressure from infectious diseases. Such extensive diversity among modern populations may indicate that geographically distinct diseases have exerted recent or perhaps on-going selection on KIR repertoires. From a practical viewpoint, this makes the choice of controls very important for all disease association studies. Additionally, it points to the possibility of finding population restricted associations of specific KIR polymorphisms with diseases.

The population HLA allele pool also plays a role in shaping the population diversity of KIR alleles and haplotypes, since most of the KIR receptors function by recognizing MHC class I ligands (Table 1.1). Different KIR receptors recognize specific motifs of HLA class I molecules. The affinity strength of those interactions is variable depending of both KIR and HLA genes and / or alleles involved [63]. KIR2DL1 and KIR2DL2/L3 recognize motifs majorly present in HLA-C alleles. KIR2DL1 binds HLA-C epitopes distinguished by the presence of a lysine in position 80 (C2 epitope or HLA-C2) while KIR2DL2 and KIR2DL3, which are alleles from the same gene, bind HLA-C and some HLA-B epitopes containing an asparagine in position 80 (C1 epitope or HLA-C1) [63].

KIR3DL1 binds HLA-B and some HLA-A possessing the Bw4 public epitope. HLA-B alleles can be categorized in HLA-Bw4 and HLA-Bw6 for containing the serological defined Bw4 and Bw6 epitopes, respectively. Those cross-reactive epitopes are defined by specific amino acid motifs in the 77-83 positions in the α1 domain of HLA-B (and some HLA-A containing Bw4 epitopes), but only Bw4 epitopes are known to be recognized by KIR3DL1 (Table 1.2) [64]. KIR3DL2 receptors bind to HLA-A*03 and HLA-A*11 allotypes, however this receptor is shows a weak inhibitory ability and an interaction dependent on the peptide bound to the HLA molecule [65].

Table 1.1: KIR molecules and their respective ligands, signalling and function [66].

KIR	Ligand	Signalling	Function
2DL1	HLA-C2Lys80	ITIM	Inhibitory
2DS1	HLA-C2Lys80 (weak)	DAP12	Activating
2DL2	HLA-C1Asn80	ITIM	Inhibitory
2DL3	HLA-C1Asn80	ITIM	Inhibitory
2DS2	HLA-C1Asn80 (weak)	DAP12	Activating
2DL4	HLA-G	?	Both
2DL5	?	ITIM	Inhibitory
2DS3	?	DAP12	Activating
2DS4	?	DAP12	Activating
2DS5	?	DAP12	Activating
3DL1	HLA-Bw4	ITIM	Inhibitory
3DS1	?	DAP12	Activating
3DL2	HLA-A (weak)	ITIM	Inhibitory
3DL3	?	ITIM	Inhibitory

While ligands for most of the inhibitory receptors are well defined, little is known about ligands for activating KIR receptors. Despite some activating KIRs having been shown to bind MHC class I ligands, those interactions are significantly weaker and their relevance is not fully understood. Distinctively, while KIR receptor interactions with ligands usually lead to either an inhibitory or activating signal, KIR2DL4 performs both functions, showing however a weak inhibitory potential [67]. It binds the non-classical HLA-G ligand in soluble form inducing activating signalling. Additionally, HLA-G expression is restricted to trophoblast cells in the foetus, suggesting its role in NK cell vascular remodelling during pregnancy [68].

Table 1.2: KIR-ligands in HLA- and HLA-B. HLA-C alleles are of the C1 or C2 group while most HLA-B alleles are Bw4 or Bw6, according to their motifs [59].

Locus	Motif	77	80
HLA-B	Bw4	N (Asparagine)	I (Isoleucine)
HLA-B	Bw4	N (Asparagine)	T (Threonine)
HLA-B	Bw4	D (Aspartic acid)	T (Threonine)
HLA-B	Bw4	S (Serine)	T (Threonine)
HLA-B	Bw6	G (Glycine)	N (Asparagine)
HLA-B	Bw6	S (Serine)	N (Asparagine)
HLA-C	C1		N (Asparagine)
HLA-C	C2		K (Lysine)

KIR associations with human pathologies

KIR genes are also frequent subjects of disease association studies due to their variability and function. Different types of pathologies such as infectious diseases, autoimmune diseases, tumour development and pregnancy complications have been associated with KIR genes, reflecting their roles in diverse physiological mechanisms involving NK cells. Those disease associations are also influenced by KIR and HLA genes epistasis, since the majority of KIR receptors function by binding to specific types of MHC class I molecules. Therefore, in most cases, an outcome associated with single KIR genes or KIR gene combinations is dependent of MHC class I genotypes, considerably increasing the complexity of these genetic investigations. Investigation of the associations between KIR genes and diseases is the focus of Chapter 2.

The resulting effect from different KIR/HLA genotypes is a consequence of the combinations of signals they are capable of generate. The balance of NK cell signals shifts towards activation or inhibition according to the individual proportion of activating or inhibitory KIR-ligand combinations. Genotypes promoting increased NK cell activation or inhibition are associated with different outcomes, being able to mutually protect or increase susceptibility against different diseases. For example, while increased activation

is associated with protection against various viral infections, it also increases the risk of autoimmune diseases, indicating there is a fine balance between efficient NK cell activation and improper activation leading to autoimmune reactions [24,69].

1.5 HLA matching in transplantation

The identification of the MHC locus resulted from early works attempting tissue transplantation in inbred mice strains, which recognized that the histocompatibility necessary to avoid transplant rejection was determined by matching donor and recipient with similar MHC [36]. From that point on, tissue transplantation developed into an established medical procedure accompanied by the necessity of MHC compatibility between donor and patient, although some organs are more affected by MHC mismatches than others. In bone marrow, and to a slightly lesser extent in kidney, transplantation knowledge of highly polymorphic HLA class I and class II genes in both donors and patients is essential for ensuring transplant success since mismatches on those genes can lead to tissue rejection through diverse pathways of the immunological system. The more similar are HLA polymorphisms from donor and patient, the better the prognostic and survival of the transplanted graft. Nevertheless, due to advances in immunosuppressive therapy some level of HLA mismatch can be accepted depending on the transplanted graft, however accompanied by undesired side effects due to its non-specific action, increasing risk of cancer and infection [70].

1.5.1 Mechanisms of transplant rejection

Rejection of solid organ transplants can be triggered by different mechanisms of the immune response and therefore vary regarding clinical manifestations and time of onset. *Acute rejection* occurs days to weeks after the transplant due mainly to the adaptive immune response from T cells recognizing alloreactive MHC molecules causing tissue damage. Since T cells have memory properties that speed up the immunological response when re-encountering antigens, any further attempts to transplant grafts from the same donor lead to *accelerated rejection*. Some individuals may have pre-existing alloantibodies against HLA or ABO group antigens due to sensitization via previous transplants, blood transfusion or pregnancy. If these are donor-specific antibodies (DSA), which can recognize alloantigens in the transplanted graft, a fast type of rejection named *hyperacute*

rejection occurs within minutes to hours. This antibody-mediated response (AMR) is characterized by reactions against vascular endothelial cells triggering blood clotting and complement cascades. *Chronic rejection* occurs months to years after the transplant, and it is characterized by gradual deterioration of the graft mediated by both cellular and humoral mechanisms (T cells and AMR). A higher number of permissible HLA mismatches is associated with shorter graft survival due to a faster progression of chronic rejection mechanisms [1].

T cell mediated allorecognition can occur directly or indirectly. Direct allorecogniton takes place when donor APCs within the graft presenting donor peptides activate recipient T cells and lead to acute rejection. Indirect allorecognition occurs when patient APCs present peptides originated from donor molecules (MHC or non-MHC), and it has been mostly associated with chronic rejection [71]. AMR is also a mechanism contributing for chronic rejection through *de novo* DSA development. Individuals with no pre-existing DSA can still receive a HLA mismatched transplant, which could then lead to post-transplant DSA development and chronic rejection even with the administration of immunosuppressive therapy [72].

HLA matching have different importance depending on the transplanted organ. Liver and corneal transplants, for example, suffer little influence from HLA mismatches. In contrast, HLA matching is crucial for kidney and even more so for bone marrow transplantation. For allogeneic hematopoietic stem cell transplantation (HSCT) from bone marrow, polymorphic HLA class I and II loci should be as close as possible in donor and recipient, having strict policies with limited degree of mismatch acceptability. HLA mismatches in this type of transplantation leads transplanted donor cells to recognize HLA antigens which are not part of the donor HLA type [73]. Despite being less strict than bone marrow transplants, HLA matching is essential for renal transplants and sensitized patients that have developed anti-HLA antibodies represent a challenge in finding suitable donors.

1.5.2 HLA matching for renal transplants

Renal transplant is recommended for patients with end-stage renal disease, being the most commonly performed type of solid organ transplant, which can be sourced from living or deceased donors [74]. For the chance of receiving a kidney from a deceased

donor, patients are allocated onto waiting lists, along with information regarding their ABO blood group, HLA class I (at least HLA-A and -B) and class II (at least HLA-DRB1) type, and the presence of anti-HLA antibodies [75]. In the UK, this data is managed by the NHS Blood and Transplant (NHSBT) recipient registry. When a potential deceased organ donor with a consent for organ donation is identified, the NHSBT receives information regarding donor's ABO blood group and HLA type. Patients in waiting lists with compatible ABO blood group and absence of anti-HLA DSA are then considered potential recipients [70].

Sensitization to HLA constitutes a major impediment for a patient to receive a kidney since pre-existing antibodies can react against donor antigens causing hyperacute rejection. Performing donor-recipient crossmatch, a test that determines if the patient has alloantibodies reacting against donor leukocytes, is crucial prior to renal transplant procedures. Time on transplant waiting lists for sensitized patients is significantly longer than for non-sensitized patients, being even more difficult for highly sensitized patients. The extent to which a recipient is sensitized can be described using calculated panel reactive antibodies (cPRA), giving the percentage of a donor panel that is reactive to a patient's serum, i.e. yields a positive crossmatch, using the donor panel HLA type as a parameter. For highly sensitized patients with >80% cPRA, finding a negative crossmatch is so challenging that they tend to accumulate on transplant waiting lists [76], being the reason why they are prioritized in registries matching ranking systems [77]. Additionally, the presence of anti-HLA antibodies in sensitized patients with mismatched transplants reduces graft survival to an extent proportional to the degree of sensitization [78].

Among other factors, ranking potential recipients in the waiting list is based on the degree of HLA mismatches particularly at the HLA-DR, -B, and -A loci, comprising from 0 to 6 possible HLA mismatches due to the codominance of those biallelic loci [79]. Lower numbers of HLA mismatches are associated with lower rates of acute rejection and longer graft survival [80]. The use of immunosuppressive therapy helps preventing acute rejection episodes and enables the transplant, despite doing little to prevent chronic rejection which affects long-term survival of the graft. Nevertheless, acute rejection still occurs in nearly 10-20% of recipients within a year of transplantation [70].

1.6 Immunogenetic databases

This section briefly presents important web-based bioinformatics databases that are used throughout this thesis, and in the case of the Allele Frequency Net Database (AFND) adapted, in various contexts in Chapter 2-4. A wider summary beyond the databases presented in is contained within [31].

1.6.1 IMGT/HLA and the Immuno Polymorphism Database (IPD)

The international ImMunoGeneTics information system® (IMGT®) [81] integrates various informatics tools related to various components of the immune system, such as genetic, proteomic, structural data from various organisms as a resulting effort from the HLA Informatics Group in collaboration with the European Bioinformatics Institute. Their sub-databases IMGT/HLA and the IPD/KIR [59] store information regarding human HLA and KIR, respectively, and were relevant for the analyses performed in the present thesis.

The IMGT/HLA database maintains reported sequences of human MHC polymorphic loci in all available resolutions, and it is the official repository for the WHO Nomenclature Committee that determines HLA allele nomenclatures for newly released sequences. Clinical HLA typing laboratories for hospitals and donor registries, commercial organizations, and large-scale genome sequencing projects are the data source for IMGT/HLA. A high standard in data quality is achieved by the strict acceptance criteria for submission of novel sequences. Similar standards are employed in the IPD/KIR database, which is the official repository for KIR polymorphisms, following standards determined by the KIR Nomenclature Committee. The IPD/KIR is part of a parent database named Immuno Polymorphisms Database (IPD), characterized by the storage of non-HLA immunogenetic data. IMGT® data can be also cross-referenced in the International Nucleotide Sequence Database Collaboration (INSDC), consisting of DNA DataBank of Japan (DDBJ) (Japan), GenBank (USA), and the EMBL-European Nucleotide Archive (ENA) (UK).

Both IMGT/HLA and IPD/KIR platforms also provide tools for several types of analysis, such as allele metadata including previous nomenclatures, multi loci sequence

alignment of HLA and KIR alleles for both nucleotide and polypeptide sequences, sequence similarity search tools (FASTA and BLAST), and bulk data download of all versions released through *ftp* in custom formats. The IMGT/HLA also provides a listing of ambiguous allele combinations which are identical over exons 2 and 3 for HLA class I and exon 2 for HLA class II.

IPD-IMGT/HLA URL: https://www.ebi.ac.uk/ipd/imgt/hla/

IPD-KIR URL: https://www.ebi.ac.uk/ipd/kir/

1.6.2 Allele Frequency Net Database (AFND)

To capture population diversity of HLA and other immune genes, the Allele Frequency Net Database (AFND) [44] was created as an online repository for the storage of immune gene frequencies in different populations across the world. AFND is co-developed by my supervisors Profs Middleton and Jones. As mentioned throughout this chapter, immunogenetic population variability influences several clinical and research fields such as associations with diseases, histocompatibility, pharmacogenetics, anthropological and evolutionary studies. Therefore, this repository provides invaluable information for a broad range of analyses in diverse research fields, containing more than a thousand populations from more than 10 million healthy individuals. AFND was initially created to store frequencies of immune gene polymorphisms, including polymorphic HLA and KIR loci, and has been in constant development with the addition of specialized data types and new tools, where some of the new developments are part of this thesis results.

AFND data is sourced from peer-reviewed publications, from populations analysed by HLA working groups and committees, populations from national registries and individual laboratories. The data to be included is verified and standardized following the criteria for data quality before being publicized in the website. These criteria include correct designation of population characteristics and their geographical locations, validation of frequency data, and compliance with official allele nomenclatures. New AFND developments and investigations as part of this thesis involved the use of HLA and KIR frequency and genotype data previously stored in the database. Those developments providing specialized tools in clinical and research fields have been

described in the latest AFND update published in the Nucleic Acids Research journal in 2015 during the course of this PhD, and are described in detail in chapters 2 and 3.

AFND URL: http://allelefrequencies.net

1.7 Research Aims

The overall aim of the thesis is to use bioinformatics approaches to understand how genetic polymorphisms in immune genes (HLA and KIR) manifest in different outcomes in important clinical scenarios. Three scenarios were examined using three different bioinformatics techniques. First, associations between KIR genes and disease risk (largely auto-immune and infectious and disease risk were examined via literature mining alongside the generation of a new public database (Chapter 2). Second, the structural polymorphism of HLA and relationship to antibody-mediated transplant rejection was investigated using adapted population genetics techniques (Chapter 3). Third, HLA associations with adverse drug reactions were examined, using a particular case study of nevirapine, and the technique of *in silico* docking (Chapter 4). Chapter 5 presents brief discussion and summarises the overall results.

Chapter 2

A database for curating the associations between killer-cell immunoglobulin-like receptors and diseases in worldwide populations

2.1 Abstract

The killer cell-immunoglobulin-like receptors (KIR) play a fundamental role in the innate immune system, through their interactions with human leukocyte antigen (HLA) molecules, leading to the modulation of activity in natural killer (NK) cells, mainly related to killing cells infected by pathogens. KIR genes are hugely polymorphic both in the number of genes an individual carry and in the number of alleles identified. This chapter describes the development of a database named KIR and Diseases Database (KDDB), capturing a large quantity of data derived from publications in which KIR genes, alleles, genotypes and/or haplotypes have been associated with infectious diseases (e.g. hepatitis C, HIV, malaria), autoimmune disorders (e.g. type I diabetes, rheumatoid arthritis), cancer and pregnancy-related complications. KDDB has been developed as part of the Allele Frequency Net Database (AFND, http://www.allelefrequencies.net), a larger platform which captures worldwide frequencies of alleles, genes and haplotypes for several immune genes, including KIR genes, in healthy populations, covering over four million individuals. KDDB has been created through an extensive manual curation effort, extracting data on more than a thousand KIR-disease records, comprising about a hundred different diseases. Furthermore, an overview analysis of KDDB showed a trend for association of activating KIR genes with autoimmune diseases. Published in two peer-reviewed journals, KDDB has been providing a new resource for understanding KIR and disease associations. Database URL: http://www.allelefrequencies.net/diseases/

2.2 Introduction

The human leukocyte antigen (HLA) and killer-cell immunoglobulin-like receptors (KIR) gene families have been implicated in many disease association studies, due to their immune function and their high variability between individuals and populations. It has been suggested that variability presented by these genes is correlated with the observation of different disease outcomes, i.e. a higher susceptibility or protection against diseases in some individuals. Differences in their genetic compositions lead to individual variability in the immune response, fine tuning its effectiveness. While most HLA genes are broadly expressed in cells and are involved in multiple pathways of the cell-mediated adaptive immune response [36,82], KIR genes are mostly expressed in natural killer (NK) cells, modulating its activation during a NK cell synapse [83].

NK cells are bone marrow derived lymphocytes that play an active role in the innate immune system by interacting with HLA class I molecules to kill pathogen infected cells [12]. Initially, NK cells were discovered as a result of their ability to target and kill tumour cell lines that expressed little or no HLA class I molecules, a mechanism defined as "missing-self" [14]. It is now known that more complex scenarios can modulate NK cell activation, which is dependent on a mixture of activating and inhibitory receptors present on the membrane and the interaction with a variety of ligands in target cells, including HLA ligands [15].

KIR receptors are the most polymorphic receptors in NK cells, and are able to contribute to its activation or inhibition. Expressed by a gene cluster in the chromosome 19, their variability relates not only to their polymorphisms (allelic variants of a single gene), but also to the different combinations of KIR genes composing haplotypes [17-21]. This variable gene and/or allele content of an individual [55], gives rise to a very large number of different genotypes that have been observed (presence/absence of KIR genes). To collect allele, haplotype and genotype frequencies of several immune genes in different healthy human populations, the Allele Frequency Net Database (AFND) was developed [84]. AFND stores large sets of data regarding HLA, KIR major histocompatibility complex class I chain related (MIC) and cytokine gene polymorphisms, and has shown to be frequently used in the immunogenetics field, being currently cited in more than 600 publications. To date, 594 different KIR genotypes in 18,981 individuals from 157 populations have been reported to AFND.

Different disease types, mainly of autoimmune and infectious nature, have been associated with KIR genetic variants. Reported results include associations with single genes (or single alleles) to associations with groups of genes and full genotypes [24,69,82,85]. A disease association is defined as a statistically significant association between a genetic element (gene, allele, genotype etc.) with a given disease outcome, either positive or negative i.e. the genetic profile makes the disease more likely/severe or less likely/severe than the control population. Due to the number of existing KIR genes and haplotypes, studies test for multiple associations with KIR varying in complexity, especially when attempting to investigate a synergic effect of combinations of genes. Furthermore, another level of complexity is added when information about HLA ligands is included. As such, the development of a database to store data regarding disease associations with those genes is a necessary step towards a more effective comprehension of such complex data.

The work described in the present chapter produced two publications in two peer-reviewed journals. The most relevant was published in 2013 by the Database journal, consisting of a detailed description of KDDB development similarly as presented in this chapter [86] (Appendix A). This publication was part of a special issue devoted to the International Society of Biocuration, and has been presented in the 6th International Biocuration Conference in the same year of publication. Contribution to this work from colleagues included as co-authors mostly consisted of data extraction from selected research papers to be included in KDDB, while planning, structuring and developing KDDB, as well as supervising the curation process, were majorly my contribution to the present work. Later, two manuscripts published in Transfusion Medicine and Hemotherapy journal (2014) [87] and in the Nucleic Acids Research journal (2015) [44] describing the latest updates in AFND include a summarized KDDB description (Appendix A). I am first author of the aforementioned publications, with a shared first authorship in the latter. KDDB development was also published (non-peer reviewed) in both European Federation for Immunogenetics (EFI) and American Society for Histocompatibility and Immunogenetics (ASHI) newsletters in 2014 (Appendix A).

This chapter describes the creation of the KIR and Disease Database (KDDB) from data curation process to database development, and demonstrates its utility to facilitate meta-analyses and trend identification by exploring whether there is evidence that the presence of activating KIR genes are significantly associated with susceptibility to auto-immune disease or protection from infectious disease or with and vice-versa.

2.3 Methods

2.3.1 Data curation

The first step towards creation of KDDB was the collection and extraction of data from peer-reviewed publications, according to the workflow shown in Figure 2.1. Published KIR and disease association studies were extracted from the HuGE Navigator (version 2.0) [88], which is a web-based tool enabling searches of the scientific literature for studies on genetic associations with diseases. The HuGE Navigator makes use of the MeSH (Medical Subject Headings) terminology, which contains standardised keywords associated with clinically-related published studies. In KDDB, we loaded MeSH terms that describe specific diseases with which associations have been found. Manual curation was performed to extract relevant data from retrieved studies. A set of consistent rules were applied to ensure that different curators extracted data in the same way (Figure 2.1). All studies identified based on the relevant MeSH terms were analysed and inserted into KDDB unless they did not fulfill one of the following criteria (also shown on Figure 2.1): (i) the article was not written in English, since we do not have the capability to translate articles at present (ii) the study design was not based on a gene frequency comparison between two samples with different clinical outcomes (future updates to KDDB will attempt to include more complex study designs), (iii) the article identified by the HuGE Navigator was not in fact related to KIR (i.e. misidentified), (iv) the study was not related to a disease specifically, but to transplantation outcomes. Studies associating transplantation outcomes and KIR have heterogeneous designs - some studies associate KIR-ligand matches/mismatches based on recipient and donor samples, and others correlate the risk of relapse with KIR combinations. These study designs are under evaluation to ascertain if they can either fit the existing KDDB schema, or will be stored in a different database schema in the future. A data validation pipeline was created to ensure that quantitative data and metadata had been correctly extracted from each publication, involving two curators reviewing the same source publication to reduce the chance for misinterpretation or copy-paste errors.

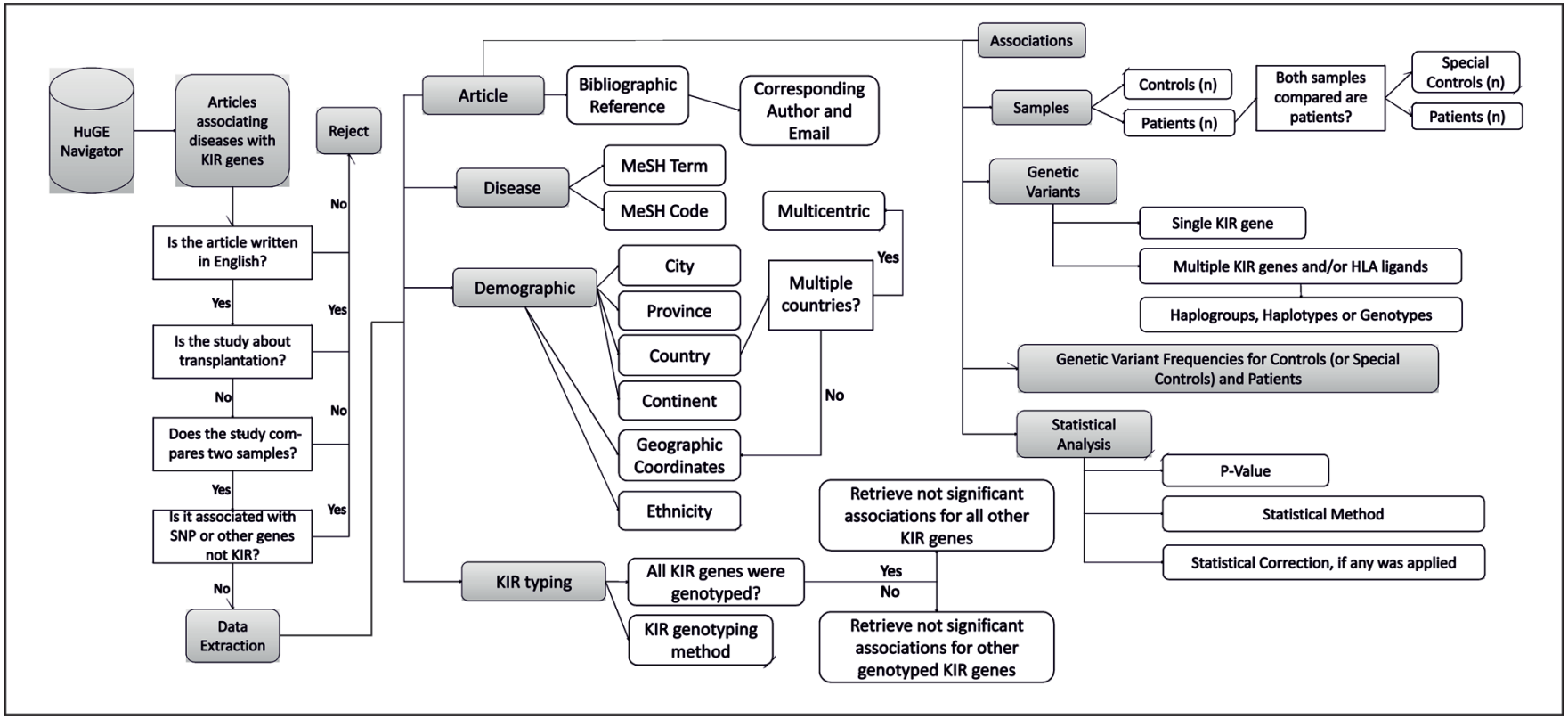


Figure 2.1: The data curation pipeline, the types of data that were extracted from each publication and the submission workflow developed within KDDB.

2.3.2 Implementation

The back-end of the database was developed using a Microsoft SQL Server relational database schema. The organization and relationship of core information stored in KDDB is described by the entity-relationship diagram presented in Figure 2.2. Three main relational entities were identified as common for all curated article: 'Study' comprises data regarding the study profile usually having single values, such as literature identifiers (PubMed ID or pmid), population data and sample sizes; 'Disease' provides additional classification of disease type and MeSH information, and it is a separate entity since a study can be associated with more than one disease classification; 'Association' comprises data regarding associations reported by individual studies; and 'Country' and 'Continent' stores standard geographic information used throughout AFND.

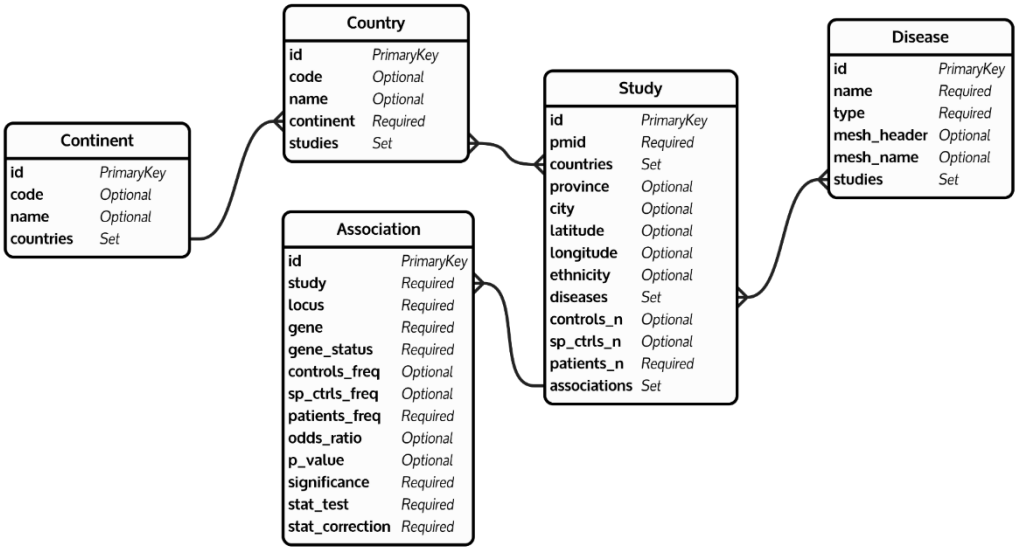


Figure 2.2: Entity-relationship diagram of KDDB schema.

Users can connect to the database using the most common web browsers. The web interface of KDDB has been created to allow users to query the database, retrieve data and submit new data sets. For that purpose, interactive web pages for querying and

submitting data were developed using the Active Server Pages (ASP) scripting environment and JavaScript language. The graphical display was designed using HyperText Markup Language (HTML) and Cascading Style Sheets (CSS), ensuring that the page will be viewable in most used web browsers. The data submission pipeline will be an important feature for future updates to KDDB, since we recognise the benefits of obtaining community input, including unpublished data sets.

In order to submit studies to KDDB, a submission form pipeline was developed which can be accessed through the AFND homepage by the menu "Submissions" and the submenu "Add KIR and disease association study" (Figure 2.3). This web form consists of four steps. The first step captures summary information about the study including the number of patients and controls. Information is also captured on the geographic location of the population, the ethnicity and the bibliographic reference. The second step captures the disease association data – the genes, alleles, haplotypes, KIR-HLA ligands, etc., the disease name, the frequency of patients and controls exhibiting the given genetic profile and the results of the statistical test. The third step (optional) allows users to upload anonymised raw data (the KIR genetic profile of every individual in the study). The fourth step allows users to review their data and submit. The submission pipeline is publicly available for users to submit their own studies, including unpublished data or studies missed in the curation process.

2.3.3 KDDB Data Analysis

Descriptive analyses of the frequency and distribution of studies associating diseases to KIR genes were performed in a subset of the associations captured by KDDB containing only: i) associations with the presence of single KIR genes (including associations with combinations of KIR and HLA ligands); ii) associations with complete information regarding the association direction (susceptibility / protection) and iii) associations with a specific disease term, excluding studies associating multiple diseases of different types (Appendix B). The exclusion of KIR combinations was implemented to avoid associations with low statistical power due to multiple comparisons, while the remaining criteria was implemented for simplicity.

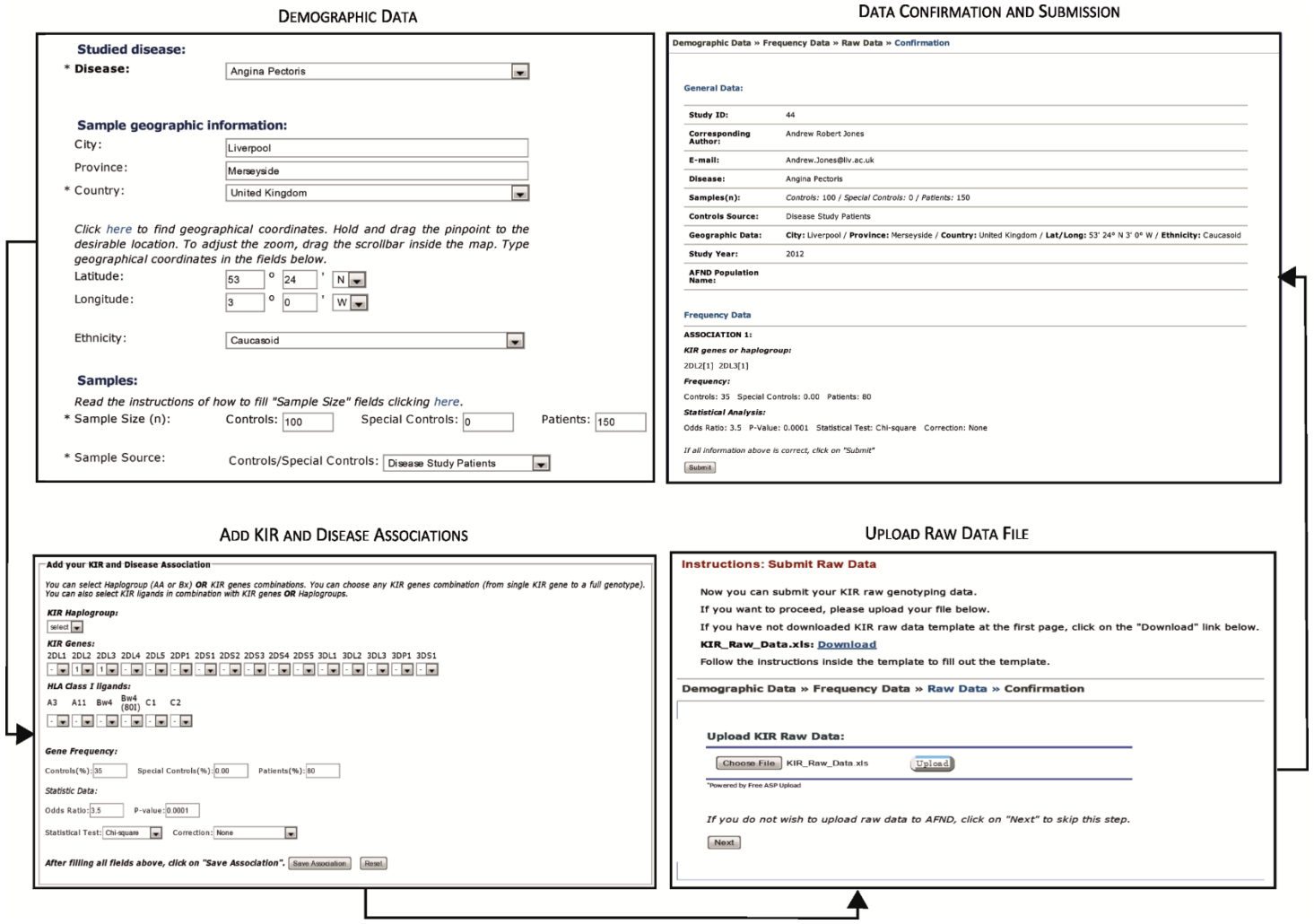


Figure 2.3: Screenshots of the data submission pipeline within KDDB. The arrows indicate the flow of the user interaction with the submission pipeline.

To investigate trends for activating or inhibitory KIR genes to be more frequently reported associated to susceptibility or protection to different disease types, the number of associations reported for KIR genes (individually or grouped by activating and inhibitory function) was compared between susceptibility and protection groups of each disease type using the Fisher's exact test based on a 2×2 contingency table. Analysis grouping by gene function excluded 2DL4 and pseudogenes. Since there are only two neurological studies, this disease type was not included in statistical analysis. Undetermined studies were also not included in descriptive analysis due to their variable aetiology. All statistical analyses were performed using R statistical packages [89] within RStudio integrated development environment (IDE) [90].

2.4 Results

2.4.1 Website organization

The KIR and Diseases Database is part of Allele Frequencies Net Database (AFND), and can be accessed through the AFND homepage (http://www.allelefrequencies.net/) using the menu "KIR" and the submenu "KIR and disease associations" (Figure 2.4) or via a direct URL access at http://www.allelefrequencies.net/diseases/, both options opening the KDDB homepage (Figure 2.5). The website interface allows the user to query and retrieve KIR and disease associations applying a collection of filters. The user can restrict the search by gene or allele, country of origin of studied samples, continent of origin of studied samples or studied disease. Those filters can be applied alone or used in combination (Figure 2.6).

Results from a query are retrieved in a table format, with each row being a different disease association with KIR (Figure 2.6). In each row, the following information is displayed: (i) row number, (ii) the associated MeSH term, (iii) the country of origin of the sample, (iv) the associated KIR profile, (v) the sample size and gene frequencies for controls and patients, (vi) odds ratio value, (vii) p-value and (viii) statistical method employed in comparisons. A link is provided, by clicking on the population name, to show the demographic information on the disease and corresponding control populations. As for normal populations in AFND, individual KIR gene frequencies or haplotype frequencies can be plotted on world maps. This enables a user to interpret

disease association risks for KIR profiles in a geographic, ethnic group or individual population-based context.

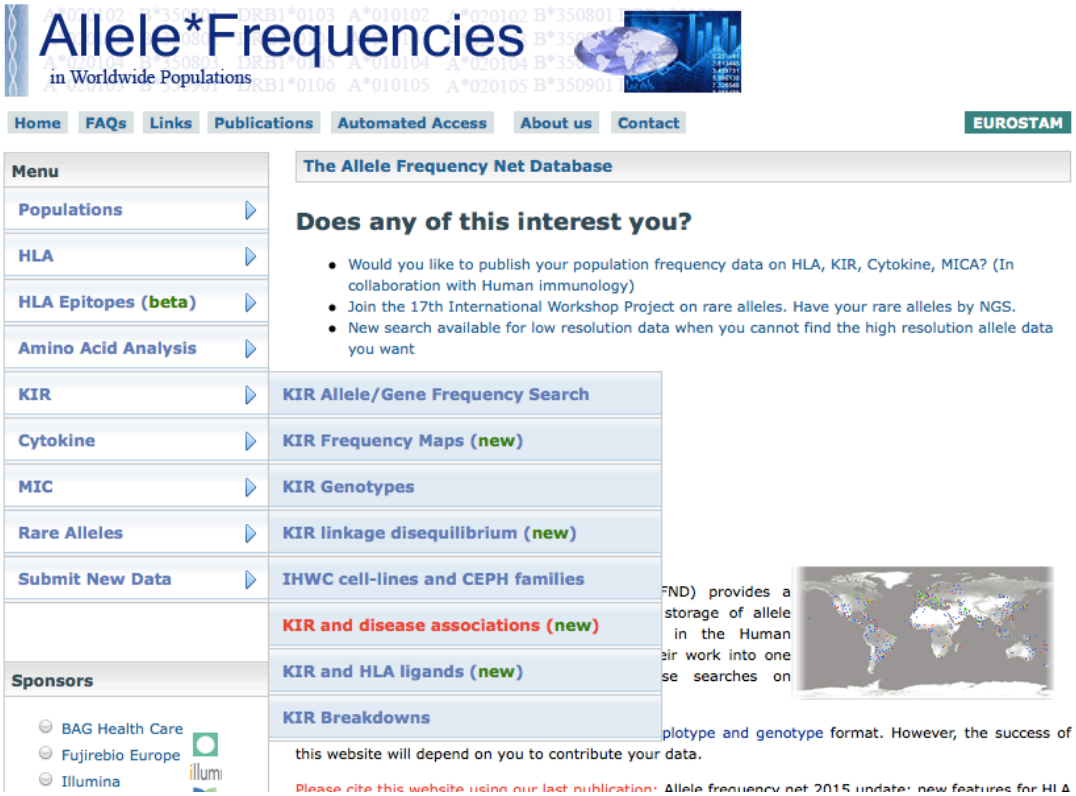


Figure 2.4: KDDB can be accessed through the AFND homepage (http://www.allelefrequencies.net/) using the menu "KIR" and the submenu "KIR and disease associations".

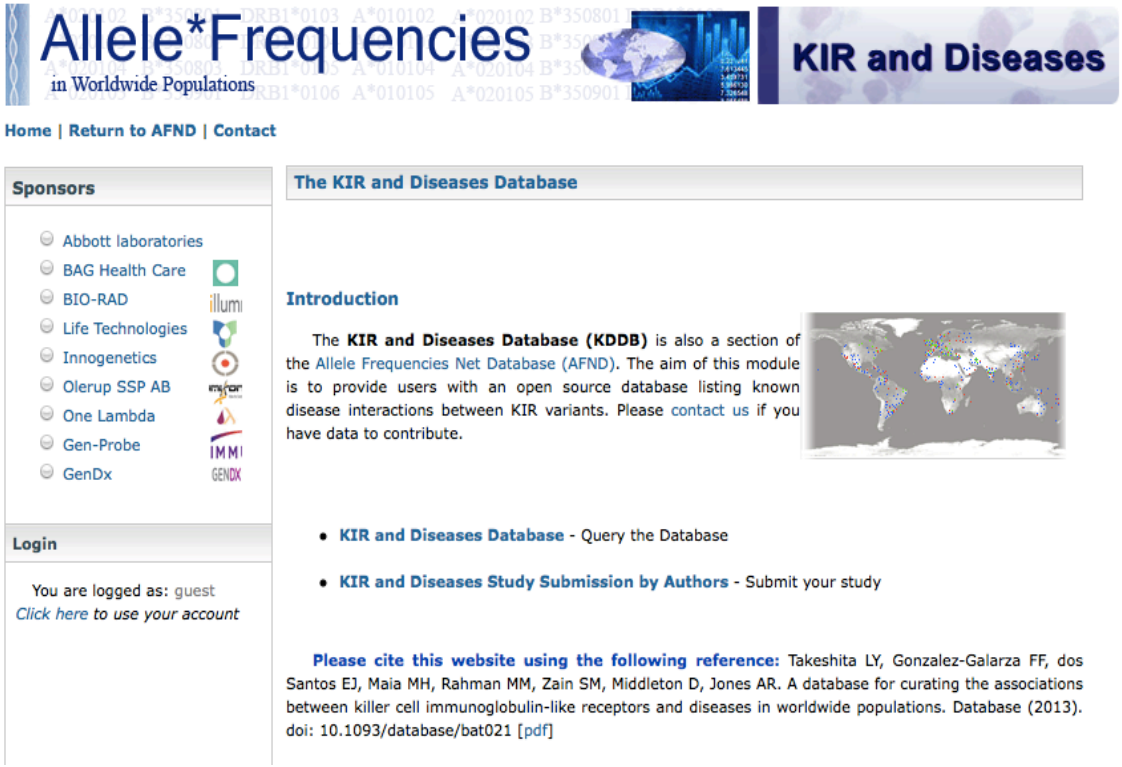


Figure 2.5: KDDB homepage providing links for querying the database and to submit new studies.

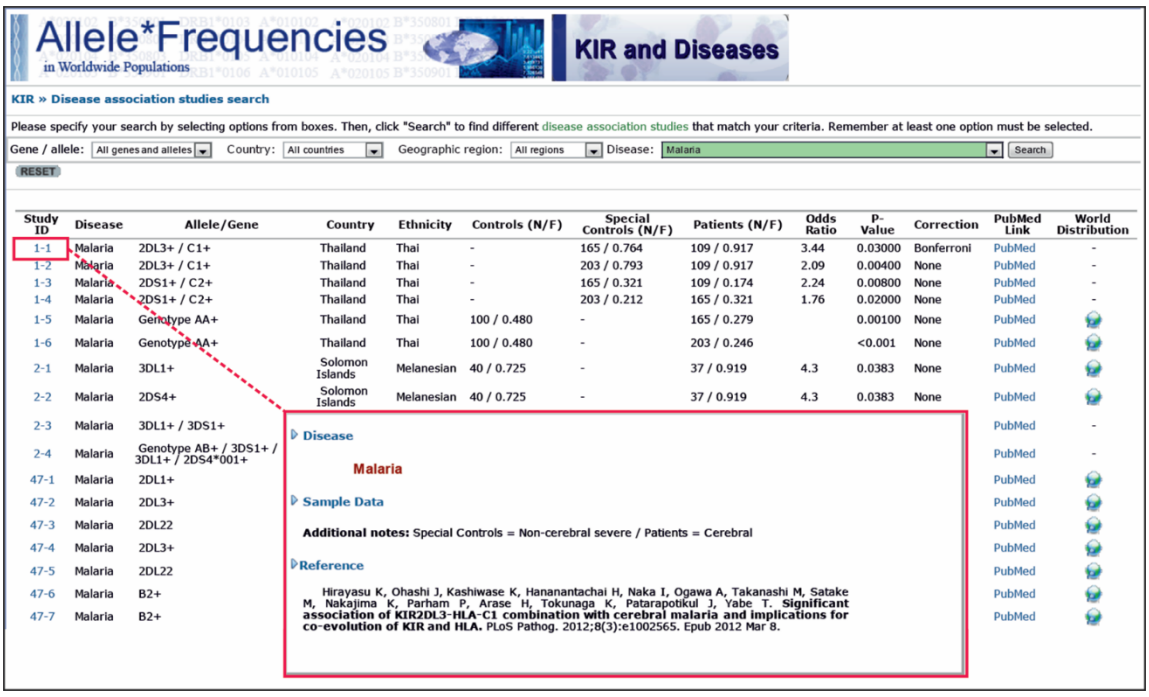


Figure 2.6: The query interface within KDDB, showing the additional detail about a given association study retrieved by following the hyperlink.

2.4.2 KDDB Content and Geographical Distribution

In the initial literature search performed alongside KDDB development, 159 articles remained after applying the exclusion criteria detailed in Figure 2.1. From all the articles, a total of 1027 KIR disease-associations were captured from 113 articles. A set of 46 articles were removed at this stage due to studies lacking mandatory data/metadata or the numerical data were inaccessible, for example displayed only on charts. The genetic associations identified in this data compilation included those with single KIR genes, profiles of combined KIR genes and / or HLA class I ligands, and full KIR genotypes. In total, 70 unique diseases have been associated with KIR across the studies in the present database. From these studies, a total of 1027 KIR records were inserted into KDDB, of which 496 are statistically significant KIR-disease associations.

After the initial curation, a total of 139 additional articles were included in KDDB by further instances of literature search and data curation, comprising a total of 427 associations, where 26 diseases distinct from the previous data collection were added, summing to a total of 96 distinct diseases investigated by studies stored in KDDB. As shown in Figure 2.7, KIR studies associated with autoimmune and infectious diseases are the majority, accounting for 34% (89) and 32% (84), respectively, followed by cancer studies with 15% (41), pregnancy complications with 8% (21) and neurological conditions with 1% (2). The remaining 10% (26) of the studies investigated diseases or conditions with undetermined caused or not fitting the previous categories (e.g. multifactorial diseases), but known to have an immune component related to its manifestation. Some studies were associated to multiple disease types, and were included in more than one disease type category.

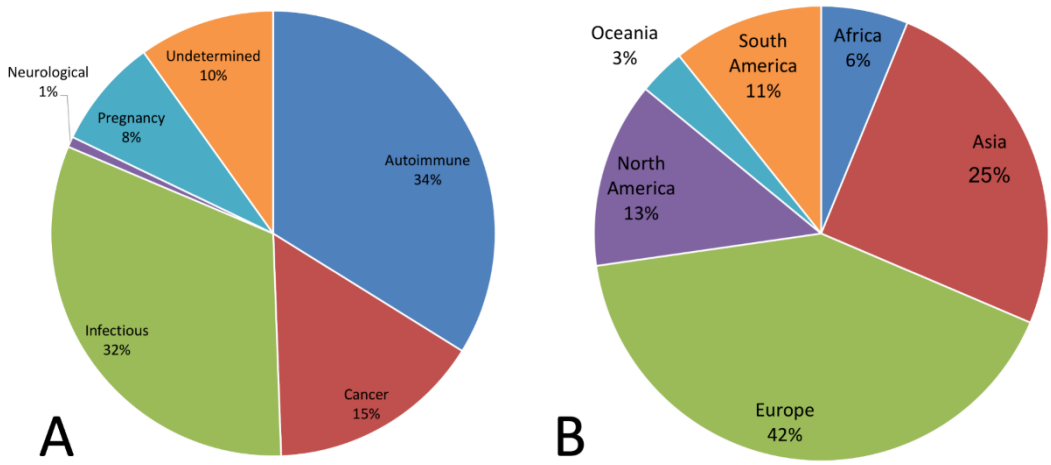


Figure 2.7: (A) Percentage of studies stored in KDDB classified by disease type. (B) Percentage of studies stored in KDDB classified by continent. Multi-centric studies or articles not clearly mentioning geographical origin of samples were not included.

KIR and disease studies have a wide geographical distribution, but most of the studies are from Europe (42%), followed by Asia (25%) and North America (13%) (Figure 2.7). Studies investigating infectious diseases have a relatively even distribution compared to autoimmune disease studies, the latter group being concentrated in certain geographical regions, being completely absent in Africa, except for Tunisia in North Africa. Some countries such as Japan, and several countries in the Eastern and Northern portion of Europe, including all Scandinavian countries, have reported autoimmune, but no infectious disease studies. Argentina, the north of Brazil in South America, and most countries in Africa and Southeast Asia have only reported KIR studies with infectious diseases. Other disease types have an even global distribution, except for the absence of cancer studies in Africa (Figures 2.8 and 2.9).

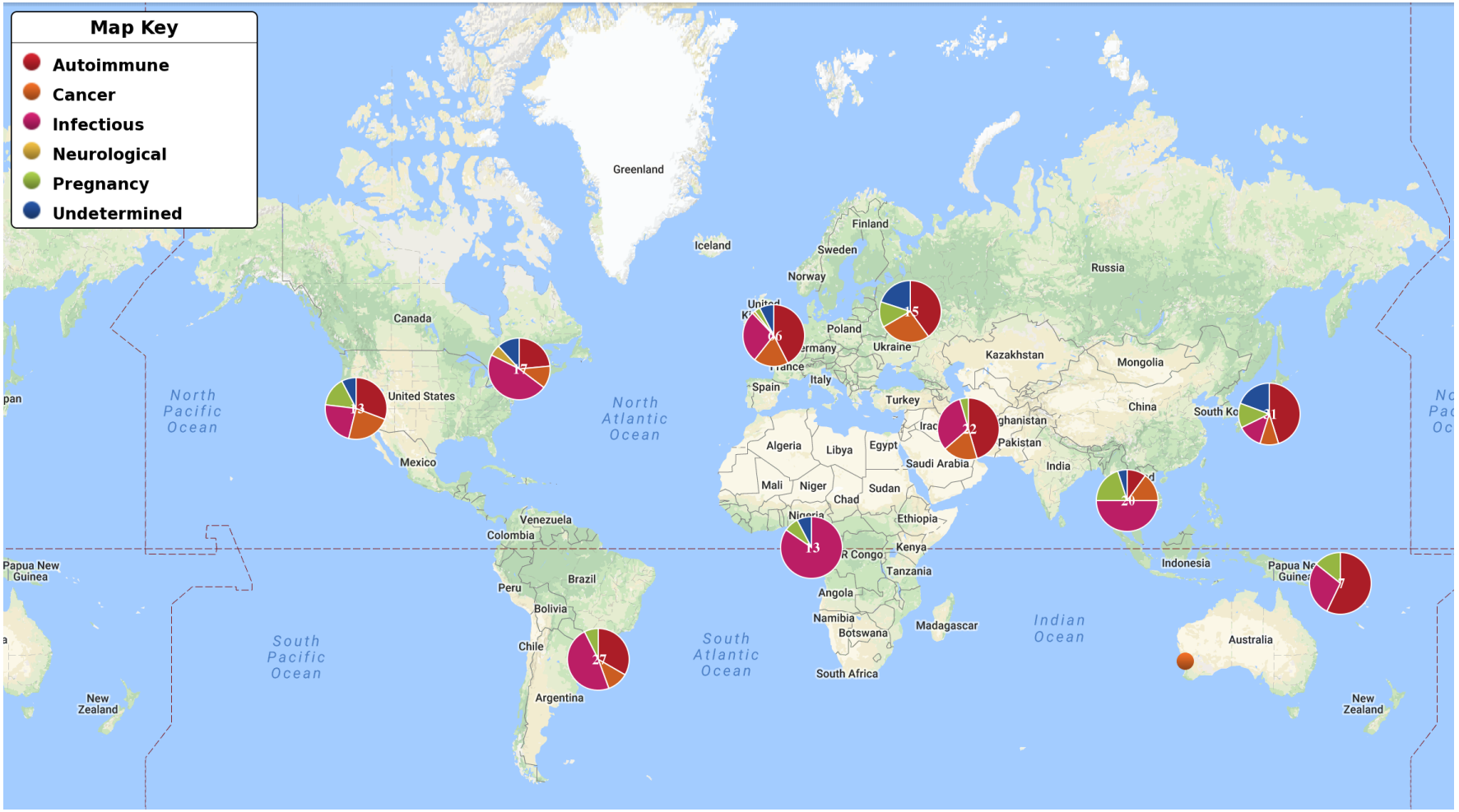


Figure 2.8: Geographical distribution of KIR studies according to the disease type investigated.

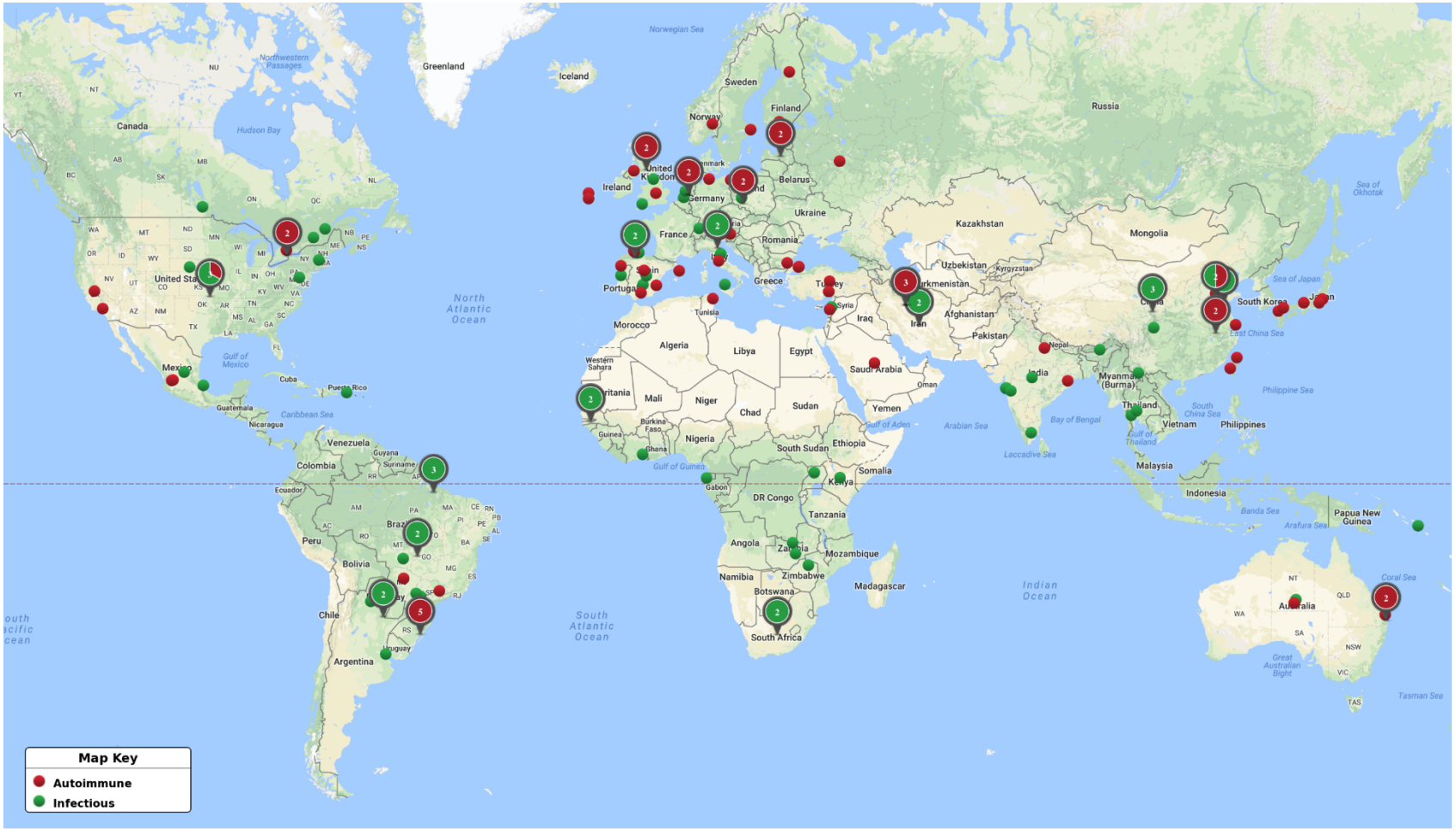


Figure 2.9: Geographical distribution of KIR studies investigating autoimmune and infectious diseases.

2.4.3 Descriptive analysis of KDDB data

The following analyses were performed using a subset of the associations present in KDDB selected according to criteria described in Methods section. From associations reported in studies, autoimmune and infectious diseases have been associated with all activating (2DS1, 2DS2, 2DS3, 2DS4, 2DS5 and 3DS1) and inhibitory genes (2DL1, 2DL2, 2DL3, 2DL5, 3DL1 and 3DL2) for both susceptibility and protection effects. For other disease types, some KIR genes have not been associated with either effect (Table 2.1), but also a lower number of studies have been published for those disease types.

Regarding the number of studies finding associations with each KIR gene, Figure 2.10 shows there is variability in the frequency of reported associations according to susceptibility and protection to disease types. Some trends are noticeable for autoimmune diseases, where 2DS1, 2DS2, 3DS1 and 2DL2 appear to have been more frequently reported associated with susceptibility than to protection to this disease type, while 3DL1 is more frequently reported associated with its protection. Furthermore, the highest number of reported associations by disease type was 2DS1 association with susceptibility to autoimmune diseases, accounting for 29% of the autoimmune disease studies included in this analysis, followed by 2DL2 in 28% of the studies (Table 2.1).

For infectious diseases, three activating KIR genes are mostly reported associated with susceptibility (2DS1, 2DS2 and 2DS3), while the other activating KIR genes (2DS4, 2DS5 and 3DS1) have been mostly associated with protection, where 2DS3 and 3DS1 show higher differences between frequencies of susceptibility and protection reported associations. Regarding other disease types, a trend for higher frequency of reported associations of 2DL1 with protection to cancer can be observed. Despite the trends observed, statistical analysis comparing the number of reported associations with each KIR gene against all other associations for susceptibility and protection groups, found only 3DL1 for autoimmune diseases and 2DL1 for cancer to be statistically significant (Table 2.1) ($P = 0.04$ and $P = 0.02$, respectively).

Table 2.1: Comparison of KIR gene associations with susceptibility and protection to diseases reported in multiple association studies found in the literature, grouped by disease type.

KIR genes	**Autoimmune (Studies = 58)**			**Cancer (Studies = 23)**			**Infectious (Studies = 51)**			**Pregnancy (Studies = 7)**			**Neurological (Studies = 2)**		
Activating	S	P	P-Value	S	P	P-Value	S	P	P-Value	S	P	P-Value	S	P	P-Value
2DS1	17 (0.29)	7 (0.12)	0.91	5 (0.22)	0 (0.0)	1.00	7 (0.14)	3 (0.06)	0.96	2 (0.29)	1 (0.14)	0.88	1 (0.5)	0 (0.0)	-
2DS2	14 (0.24)	7 (0.12)	0.79	3 (0.13)	2 (0.09)	0.78	6 (0.12)	4 (0.08)	0.86	3 (0.43)	0 (0.0)	1.00	1 (0.5)	0 (0.0)	-
2DS3	4 (0.07)	4 (0.07)	0.38	3 (0.13)	1 (0.04)	0.93	9 (0.18)	3 (0.06)	0.99	1 (0.14)	0 (0.0)	1.00	0 (0.0)	0 (0.0)	-
2DS4	5 (0.09)	1 (0.02)	0.95	2 (0.09)	2 (0.09)	0.64	3 (0.06)	4 (0.08)	0.53	0 (0.0)	2 (0.29)	0.21	1 (0.5)	0 (0.0)	-
2DS5	3 (0.05)	3 (0.05)	0.44	3 (0.13)	0 (0.0)	1.00	4 (0.08)	6 (0.12)	0.40	2 (0.29)	1 (0.14)	0.88	1 (0.5)	0 (0.0)	-
3DS1	11 (0.19)	3 (0.05)	0.96	4 (0.17)	2 (0.09)	0.87	3 (0.06)	8 (0.16)	0.12	2 (0.29)	0 (0.0)	1.00	1 (0.5)	0 (0.0)	-
Total Activating	54	25	**0.04***	20	7	**0.004***	32	28	0.21	10	4	**0.04***	5	0	-
Inhibitory															
2DL1	8 (0.14)	7 (0.12)	0.36	1 (0.04)	7 (0.3)	**0.02***	2 (0.04)	2 (0.04)	0.71	0 (0.0)	2 (0.29)	0.21	0 (0.0)	0 (0.0)	-
2DL2	16 (0.28)	10 (0.17)	0.61	4 (0.17)	4 (0.17)	0.57	7 (0.14)	10 (0.2)	0.34	0 (0.0)	1 (0.14)	0.48	0 (0.0)	0 (0.0)	-
2DL3	7 (0.12)	7 (0.12)	0.28	2 (0.09)	4 (0.17)	0.27	9 (0.18)	14 (0.27)	0.21	0 (0.0)	0 (0.0)	1.00	0 (0.0)	0 (0.0)	-
2DL5	8 (0.14)	5 (0.09)	0.63	1 (0.04)	2 (0.09)	0.45	5 (0.1)	3 (0.06)	0.88	1 (0.14)	1 (0.14)	0.74	1 (0.5)	0 (0.0)	-
3DL1	4 (0.07)	8 (0.14)	**0.04***	4 (0.17)	4 (0.17)	0.57	4 (0.08)	5 (0.1)	0.53	0 (0.0)	1 (0.14)	0.48	1 (0.5)	2 (1)	-
3DL2	1 (0.02)	1 (0.02)	0.63	0 (0.0)	0 (0.0)	1.00	1 (0.02)	1 (0.02)	0.76	0 (0.0)	0 (0.0)	1.00	0 (0.0)	0 (0.0)	-
Total Inhibitory	44	38	-	12	21	-	28	35	-	1	5	-	2	2	-
2DL4	0 (0.0)	0 (0.0)	1.00	0 (0.0)	0 (0.0)	1.00	0 (0.0)	0 (0.0)	1.00	0 (0.0)	1 (0.14)	0.48	0 (0.0)	0 (0.0)	-
2DP1	1 (0.02)	3 (0.05)	1.00	0 (0.0)	0 (0.0)	1.00	0 (0.0)	0 (0.0)	1.00	0 (0.0)	1 (0.14)	1.00	0 (0.0)	0 (0.0)	-
3DP1	2 (0.03)	0 (0.0)	1.00	0 (0.0)	0 (0.0)	1.00	0 (0.0)	0 (0.0)	1.00	0 (0.0)	0 (0.0)	1.00	0 (0.0)	0 (0.0)	-

'Studies' = Number of studies found in the literature for the respective disease type. 'S' = Susceptibility; 'P' = Protection. 'Pregnancy' refers to pregnancy complications. The frequency reported between parenthesis is the number of associations reported divided by 'Studies'. 2DL4 has been shown to have both functions (Inhibitory / Activating), and 2DP1 and 3DP1 are pseudogenes. * P-values < 0.05, i.e. statistically significant differences between the number of susceptibility and protection KIR gene associations found in the literature related to different disease types. Statistical analysis was not performed for 'Neurological' studies due to the low number of studies available.

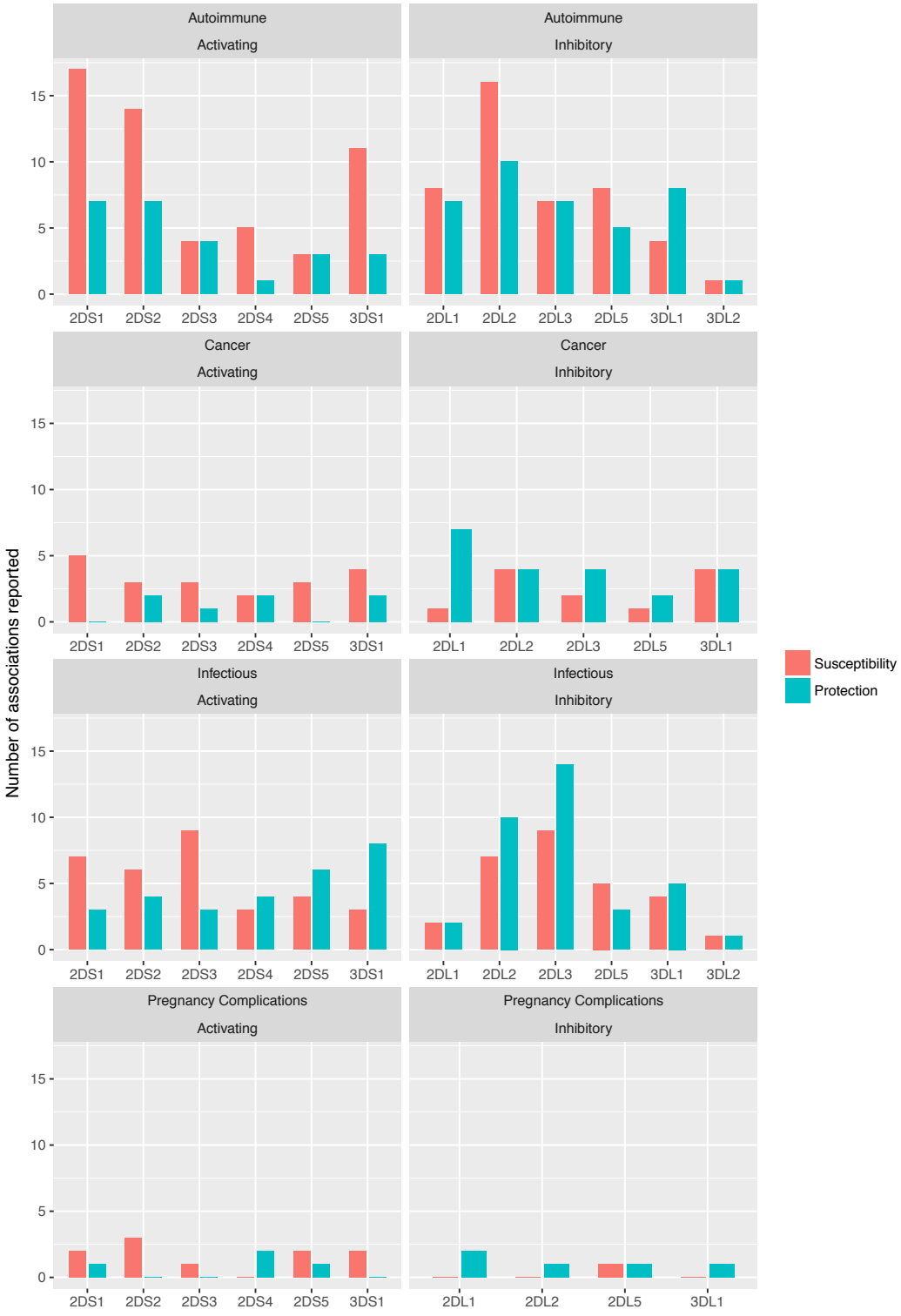


Figure 2.10: Distribution of individual KIR gene associations with susceptibility or protection to disease types. The y-axis is the count of the number of studies reporting a statistically significant association of the given gene with a given disease belonging to one of the disease classifications, either making individuals more susceptible (green) or more protected (red) under the four general headings. The KIR genes have been divided into those understood to be "activating" (left panel) and "inhibitory" (right panel).

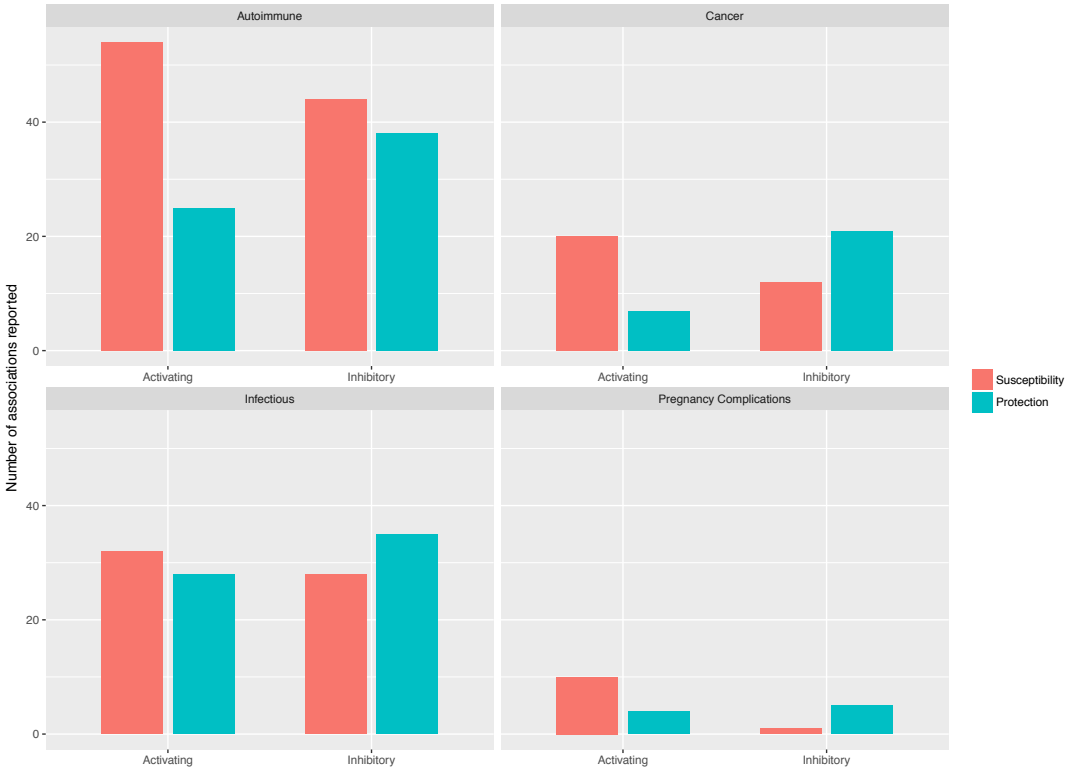


Figure 2.11: Distribution of KIR genes associations with susceptibility or protection to disease types according to their function. The y-axis is the count of the number of studies reporting a statistically significant association of an activating or inhibitory KIR gene with a given disease belonging to one of the disease classifications, either making individuals more susceptible (green) or more protected (red) under the four general headings.

Grouping genes by function also shows variability in the frequency of reported KIR gene associations with susceptibility and protection to disease types (Figure 2.11). For autoimmune diseases, cancer and pregnancy complications, enrichment of activating KIR gene associations with susceptibility to these disease classifications can be observed. An opposite trend of inhibitory KIR gene associations with protection is observed only in relation to cancer and pregnancy complications, but not for autoimmune diseases. For infectious diseases, only a slight difference is observable between the compared groups. Statistical analysis comparing the number of reported associations with KIR genes grouped by function between susceptibility and protection groups, found enrichment of activating KIR genes in susceptibility associations to autoimmune diseases, cancer and pregnancy complications to be statistically significant (Table 2.1, $P = 0.04$, $P = 0.004$ and $P = 0.04$, respectively).

For a detailed analysis of the diseases involved in the previously mentioned analysis, an overview of the proportions of activating KIR gene associations with susceptibility to individual diseases belonging to autoimmune, cancer and pregnancy complications disease classifications is shown in Figure 2.12, revealing a wide range of autoimmune diseases and cancer types covered by those disease types. Diabetes mellitus type 1 accounts for the larger proportion of the associations of activating KIRs with susceptibility to autoimmune diseases (approximately 26%), while acute lymphoblastic leukaemia and colorectal neoplasms accounts for the larger proportion of the associations of activating KIRs with susceptibility to tumours (45% and 25%, respectively). In contrast, associations of those genes with susceptibility to pregnancy complications are all related to recurrent (70%) and spontaneous (30%) miscarriage.

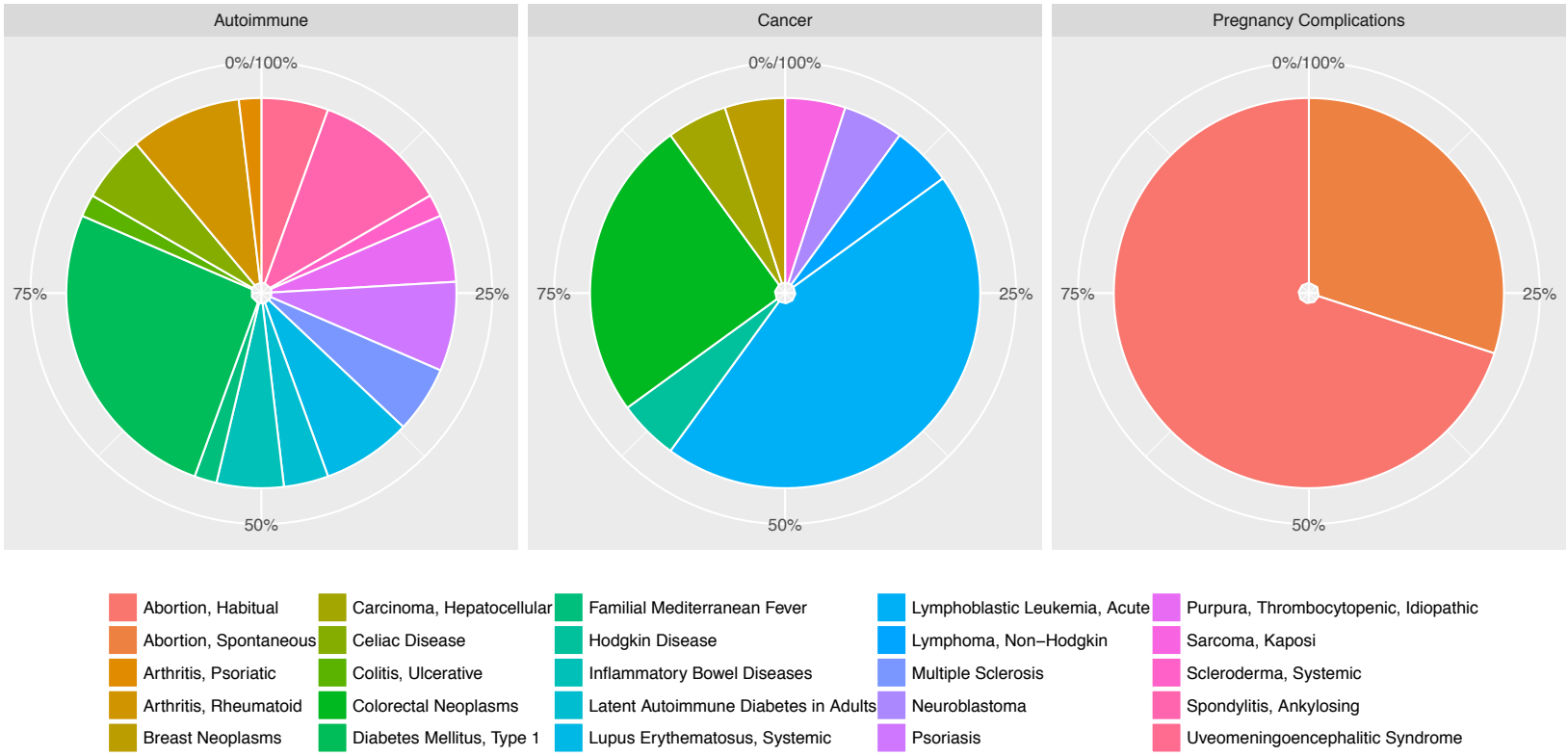


Figure 2.12: Proportion of reported KIR associations associated with susceptibility to diseases belonging to autoimmune, cancer and pregnancy complications classifications of disease types.

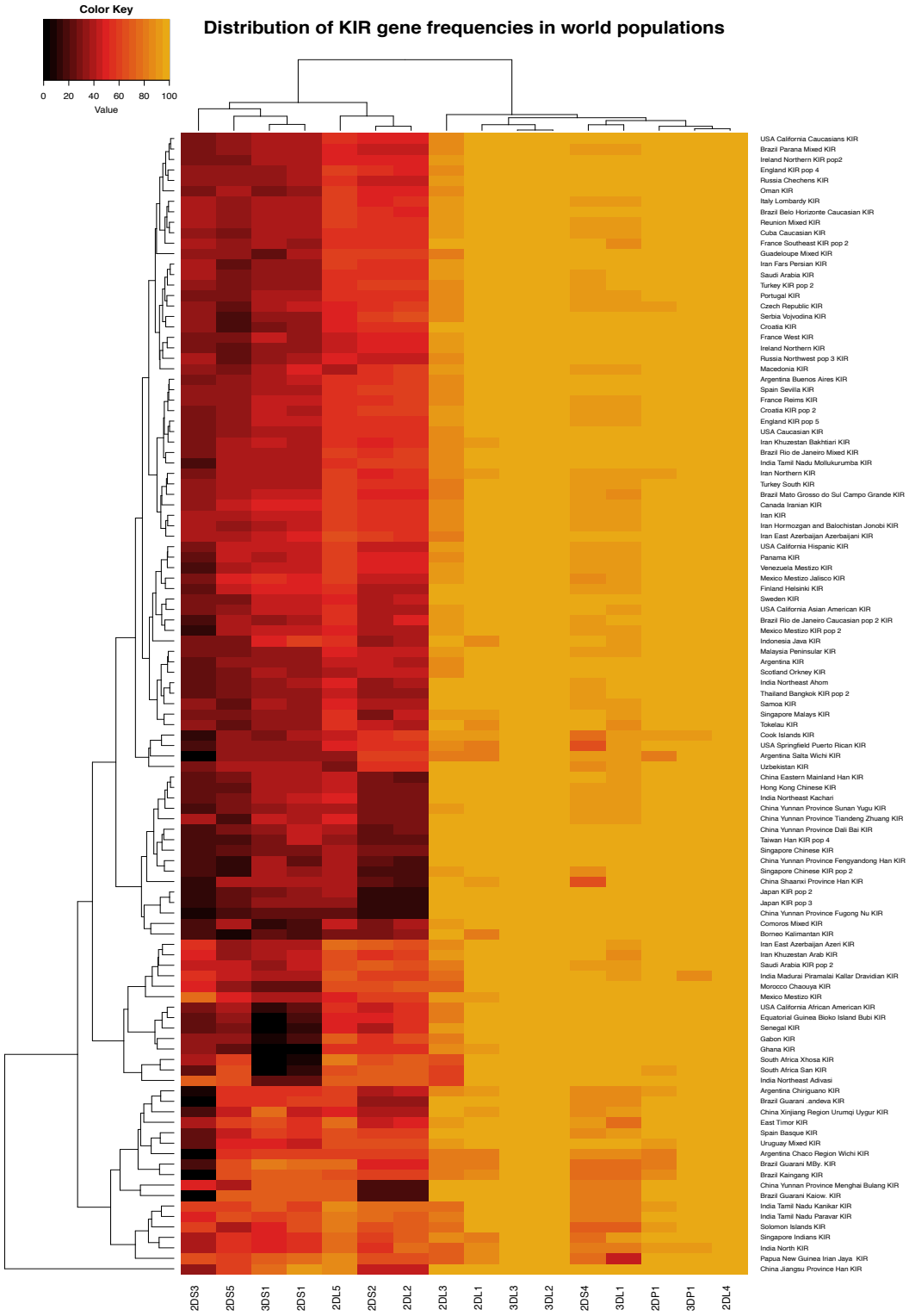


Figure 2.13: Clustered heatmap of KIR gene frequencies across populations in AFND. KIR genes belonging to A haplotypes (major branch on the right side of the x-axis) show higher frequencies in populations comprising mostly inhibitory KIR genes, while KIR genes exclusive to B haplotypes (major branch at the left side of the x-axis) show lower and most variable frequencies, containing mostly activating KIR genes. The closer KIR genes are clustered, the higher the chance they have strong linkage disequilibrium.

Since certain KIR genes have strong linkage disequilibrium, a clustered heatmap was generated for KIR gene frequencies across populations present in AFND to investigate KIR genes clustered together that can be influencing disease associations with KIR genes (Figure 2.13). KIR genes from the less variable A haplotype show generally higher frequencies in populations and contain mostly inhibitory KIR genes, while KIR genes from B haplotypes show lower and most variable frequencies, containing mostly activating KIR genes. Among B haplotype genes, strong linkage disequilibrium between two centromeric KIR genes (2DS2 and 2DL2) and between two telomeric KIR genes (2DS1 and 3DS1) are evidenced by their clustering, showing very similar gene frequencies across populations.

2.5 Discussion

An overview of the studies included in KDDB identified about a hundred distinct diseases of diverse nature associated with KIR genes, suggesting NK cell involvement in underlying mechanisms of several biological pathways. They have a wide world distribution, however most of the studies originate from developed countries. Most of the disease types associated with KIR seem to have an even geographical distribution, except for autoimmune diseases. This differential distribution of KIR studies with autoimmune diseases is in accordance with the 'hygiene hypothesis', stating that a lower incidence of infectious diseases in a population, due to economic development or climate, is correlated with the increasing incidence of autoimmune and allergic diseases [91]. However, differences may also be due to socio-economic factors in developing countries affecting access to health care and research development [92,93].

An individual NK cell simultaneously expresses distinct KIR receptors according to the individual KIR genotype, where the resulting profile contributes to the modulation of NK cell activity. In summary, the characteristics from each individual KIR genes that may affect their combined effect on NK cell activation are: i) KIR function, which can be activating, inhibitory or both, ii) their ligand specificity and iii) structural differences affecting binding to ligands, existing between KIR receptors with similar function, or between polymorphisms of a specific KIR gene.

Utilizing disease association studies within KDDB, an explorative analysis of reported statistically significant KIR and disease associations was performed, focusing on the

investigation of trends related to KIR gene function for different disease type classifications (autoimmune diseases, infectious diseases, cancer and pregnancy complications) and comparison between the association effect (susceptibility or protection). Reported KIR gene associations with susceptibility to autoimmune diseases, cancer and pregnancy complications show enrichment of activating KIRs, where more studies have reported association between susceptibility to those disease types and activating genes than one would expect by chance (Table 2.1). For autoimmune diseases and cancer, those associations cover a range of disorders, while only miscarriage (habitual and spontaneous) is covered by the pregnancy complications category.

Nevertheless, only miscarriage and pre-eclampsia have been associated with KIR genes, and the enrichment of activating KIR associations with susceptibility to miscarriage found in the present study suggests a potential role of these genes in miscarriage occurrence, but not in pre-eclampsia. Uterine NK (uNK) have a specialized function during trophoblast invasion in pregnancy, however the mechanisms involving activation of uNK cells are not fully understood [23]. It has been shown that both excessive activation and inhibition of uNK cells interacting with trophoblast cells correlate with birth weight and foetus mortality [94]. Additionally, despite the potential role of KIR2DL4 recognizing HLA-G ligand with restricted expression in foetal trophoblast cells, only one study in KDDB found an association of this gene with pre-eclampsia. It should be noted, however, that the number of studies investigating KIR and pregnancy complications in the present analysis is limited.

It has been suggested that KIR profiles leading to lower inhibition and higher activation would be beneficial in infections and at the same time constitute a risk to autoimmunity and cancers that have an inflammatory component [24]. Present results support the hypothesis of a relationship between enrichment of activating KIR gene associations and susceptibility to autoimmune and cancer, but did not indicate clear relationships with protection to infectious diseases.

Most of the studies associating KIR and infectious diseases involve 2DL2 (20%) and 2DL3 (27%) conferring protection effect, but the frequency of studies associating the same genes with susceptibility to infectious diseases are similar (18% and 14%, respectively). A possible explanation for those inhibitory genes to be associated with both susceptibility and protection against infections could be due to their function in recognizing HLA ligands. Interactions between KIR and HLA-C ligands C1 (recognized

by 2DL2 and 2DL3) and C2 (recognized by 2DL1) appear to be the dominant mechanisms controlling NK cell activation [95]. These receptors have also been shown to recognize peptides in HLA ligands, although with a much broader specificity compared to TCR receptors from T cells [95]. While 2DL1 shows a consistent high frequency across populations, 2DL2 and 2DL3 frequencies are relatively more variable (Figure 2.13), being more likely for them to be contributing to the variability of NK cell response to infections.

The data present in KDDB could in theory also be used in more robust systematic review and meta-analysis methods specific to observational studies. A systematic review consists of locating individual studies investigating a common research question, while reviewing their methods and results regarding their eligibility and methodological quality. A meta-analysis consists of a statistical analysis of the resulting data typically derived from a systematic review, generating a new result from the overall pooled dataset. Additionally, meta-analysis methods include assessment of publication bias and heterogeneity among studies (i.e. different magnitudes or different directions across results) [96].

However, there are limitations to the application of meta-analysis methods in KIR and disease studies due to the variability and complexity of KIR genes. Genetic variability among populations acts a confounder factor in meta-analyses, requiring stratification of the reviewed pooled dataset by ethnicity or application of multivariate analysis [97,98]. Furthermore, combinations of multiple KIR genes within haplotypes interacting with variable ligands could influence disease outcome, whereas most KIR studies compare only individual KIR genes between case and control groups. KIR and disease meta-analyses that have been published to date investigate the effect of individual KIR genes in specific diseases [99-104]. These studies also stratify populations by ethnicity, using broad ethnical classifications, such as 'Asian', 'Caucasian' and 'Black', which have been the subject of debate in the immunogenetic field [105]. The data in KDDB, even when examining one particular disease, tends to split across different ethnic groups, which should not be pooled for a meta-analysis. As such, traditional meta-analysis methods might be appropriate for a selection of diseases covered in KDDB, for which there are multiple studies on the same ethnic group e.g. type I diabetes, but these methods would have to be used with caution if there are doubts that other environmental factors were not well matched in different studies. Additional studies investigating combinations of KIR genes and their interactions with HLA ligands may be necessary to perform meta-analyses capable of covering the effect of multiple KIR and HLA genes in diseases, beyond what is currently captured by KDDB.

The first release of KDDB included only data identified by the HuGE Navigator. Since HuGE Navigator does not retrieve all studies, additional search strategies, for example via Pubmed and Web of Knowledge, have been employed to locate studies missed in the first pass curation process. Studies that do not fit into the simple model of a case-control disease association study have been excluded from the curation process. Capturing more complex stratification studies is possible in KDDB, but will necessitate either some loss of granularity of the data, or the development of a much more complex schema and display interface. Another limitation of KDDB is that many disease studies, especially those that do not find statistically significant associations, are not published and there is a risk that resources such as KDDB could suffer from publication bias. The web form for data submission allows researchers to submit their data, even if it is unpublished or does not contain a statistically significant association.

Other tools in AFND also allow uploading of anonymised raw data (individual KIR type and HLA ligands) to enable improved quality control measures (such as validation of frequency calculations) and to enable advanced analyses of the data. For example, having the individual data available will allow analyses such as looking at disease associations in the centromeric or the telomeric regions and evaluating the influence of LD in different populations.

2.6 Conclusions

Over the last ten years of existence, AFND has provided the immunogenetics and histocompatibility community with an online repository for the examination of immune gene frequencies in different healthy populations. Development of KDDB aims to cover disease studies that have been associated with KIR genes and to include studies in which no significant association has been found, to avoid publication bias. Analysis of KDDB content provided insights on the relationship of KIR functionality with different disease classifications, showing an enrichment of activating KIR genes reported associated with susceptibility to autoimmune diseases, tumour development and miscarriage. Future work regarding database development will consist of improvements to KDDB interface to facilitate data retrieval and database reorganization for accommodation of study types not fitting the current schema. Additional data in other sections of AFND connecting to data in KDDB are also envisioned, since KIR gene frequencies, genotyping data and LD data are important assets for investigating KIR associations. Finally, KDDB facilitates meta-

analyses and data re-use to understand the underlying function of KIR genes in a variety of disease processes, and has the potential to fill the existing gap in literature from publication bias.

Chapter 3

HLA epitope matching in world populations

3.1 Abstract

The presence of anti-HLA donor-specific antibodies (DSA) in transplant patients is a crucial factor related to rejection. These antibodies target specific regions of HLA proteins that are different from the transplant patient's HLA proteins – referred to as HLA epitopes. Current efforts in HLA matching for transplantation often attempt to minimise the number of allele-level mismatches. Each allele is distinct from others based on any non-synonimous variation considering the whole DNA sequence at any position. However, HLA comparison on a sequence level does not reflect protein structure similarity as an immunological target, since individuals with different HLA genotypes may be able to carry the same set of HLA epitopes. A matching strategy based solely on HLA allele nomenclature ignores those cases, even though they may not elicit an anti-HLA DSA response. Similar to allele frequencies, understanding HLA epitope frequencies in world populations could be important information for estimating the probability of finding a suitable donor in a given geographical region or human population. The first aim of this chapter is to describe the development of the HLA Epitope Frequency Database (HLA-EpiDB) as a new feature of AFND, being publicly accessible for users. HLA-EpiDB was generated by mapping HLA epitopes definitions to genotype/haplotype data stored in AFND to generate population-level HLA epitope frequencies. The second aim of this chapter is to present an explorative analysis of the use of HLA epitope definitions in alloreactivity profile matching using single antigen bead (SAB) profiles from sensitized patients. Comparisons between alloreactivity profile matching using HLA epitopes and high resolution HLA allele matching reveals that differences in matching proportions obtained using both methods are influenced by current limitations in both SAB assays and HLA epitope definitions. Despite these analyses suggesting that optimization of alloreactivity matching requires improvements in

both methods, it also highlights a potential use of HLA epitope definitions in increasing the granularity from SAB assay results. A better understanding of HLA epitopes can contribute to reducing the chance of positive crossmatch in sensitised patients, while knowledge of HLA epitope frequencies help transplant clinicians and the immunogenetics research community to better understand and determine HLA mismatch acceptability across worldwide populations. Database URL: http://allelefrequencies.net/hlaepitopes/

3.2 Introduction

Early transplantation studies using tumours in mice led to the establishment of the tissue compatibility concept, where the major histocompatibility complex (MHC) was recognized as a genetic region containing genes to be matched prior to transplantation [106]. Within the MHC region, the HLA gene complex contains the most relevant genes for tissue compatibility due to their extensive variability and broad expression in tissues, as differences in HLA gene type between donor and recipient can lead to rejection of the transplanted tissue triggered by an immune response against non-self molecules. HLA class I and class II molecules play a key role in the adaptive immune response, being responsible for antigen presentation to cytotoxic (CD8+) and helper (CD4+) T cells, respectively. While HLA class I molecules are ubiquitous expressed in cells, HLA class II expression is mostly restricted to APCs [36].

"Classical" HLA class I (HLA-A, -B and -C) and class II (HLA-DRB, HLA-DP and HLA-DQ) are the most polymorphic HLA loci, and therefore considerably important in transplantation. According to IMGT/HLA database release 3.29.0.1, 12,544 alleles have been described for "classical" HLA class I loci and 4,622 alleles have been described for "classical" HLA class II [107]. Most of the variation between alleles is observed within the peptide-binding site (or antigen recognition site) region, as selective pressures from a pathogen-rich environment favouring the maintenance of diverse HLA molecules that can present multiple peptide "shapes" [41].

However, all this necessary variability for defending the organism against pathogens represents a problem in the unnatural transplantation scenario. A transplant of a foreign tissue (allograft) containing molecules not present in the recipient triggers cellular and humoral alloimmune responses targeting non-self molecules present in the allograft. The

extent of HLA mismatching between donor and recipient is correlated with the risk of rejection and graft loss, but HLA matching requirements vary according to different transplant types [108]. Haematopoietic stem cell transplantation (HSCT) has the strictest matching requirements, needing high resolution (two field allele nomenclature) matching for most 'classical' HLA loci [109]. For logistical purposes, renal transplants only use low resolution (one field allele nomenclature) matching for HLA-A, -B and -DR loci and absence of anti-HLA DSA [73] (see Chapter 1 for more details on HLA allele resolution levels). The presence of anti-HLA DSA pre-transplant is associated with hyperacute rejection – an immediate immune reaction against the transplanted organ – while DSA formed post-transplant are associated with acute and chronic antibody-mediated rejection (AMR) – characterized by acute or chronic tissue injury with microvascular inflammation [108,110,111].

3.2.1 Anti-HLA Alloantibody Detection

A patient with anti-HLA DSA, or a 'sensitized' patient, has had previous contact with non-self HLA molecules through previous transplants, pregnancy or blood transfusion [108]. Some patients are considered 'highly sensitized', having developed antibodies against most HLA antigens, being extremely challenging for them to find a suitable donor [112]. The level of sensitization in patients is defined by the PRA measure, consisting in the proportion of the population to which a patient will react due to pre-existing alloantibodies. In 2006, more than 15,000 patients on the waiting list for renal transplantation in the United States were reported sensitized (defined as PRA > 20%) and nearly 50% of those patients had a PRA > 80% [113].

Detection of those alloantibodies was initially performed using cell-based assays containing lymphocyte cell panels, but introduction of solid-phase assays using purified or recombinant HLA class I and class II molecules improved the sensitivity and specificity of anti-HLA antibody detection [114]. A commonly used solid-phase assay uses multiplexed single antigen beads (SAB) coated with single HLA antigens and flow cytometry technology to detect binding of human IgG antibody (Luminex®) [115]. SAB assays can detect alloantibodies against class I (HLA-A, -B and -C) and class II molecules (HLA-DR, HLA-DQ and HLA-DP), with different degrees of resolution and allele coverage.

SAB assay results report the mean fluorescence intensity (MFI) which is a semi-quantitative estimation of the antibody level. Despite also being used to monitor antibody levels in patients especially post-transplant [116], SAB results are mainly interpreted qualitatively through the definition of an MFI cut-off value. Beads containing HLA types showing MFI values above the cut-off value are considered positive for the presence of alloantibodies against the HLA antigens present on a giving bead. This cut-off value varies across transplant laboratories since there is no consensus regarding adequate thresholds. Therefore, laboratories define the cut-off value based on the MFI levels observed in relevant controls and laboratory experience from clinical results obtained over time, falling usually in the range of 1000-2000 MFI [117].

Ultimately, interpretation of MFI values from SAB assays should be analysed individually in conjunction with patient history. MFI values can be affected by several factors related to the SAB assay, to the patient serum, or to characteristics specific to some HLA epitopes recognized by patient's antibodies, such as epitope sharing, affecting its antigenicity. Different HLA types on different beads share common HLA epitopes leading a specific anti-HLA antibody to bind on multiple beads resulting in reduction of the MFI levels reported for single beads [118].

In the transplant scenario, the humoral immune system sees each HLA type as a collection of HLA epitopes capable of being recognized by alloantibodies, where mismatched epitopes between HLA types from donor and patient result in antibody response. This rationale is leading clinicians and researchers to look at HLA matching from a structural perspective, using HLA epitopes as units to be matched.

3.2.2 HLA Epitopes and Structural Matching

In solid organ transplantation, the production of anti-HLA DSA will occur due to B cell recognition of non-self HLA epitopes. The HLA molecule surface with its multiple polymorphic amino acids represents a collection of possible HLA epitopes to elicit antibody production, if mismatched. Despite laboratory identification of several HLA epitopes [119], current experimental findings depend on the limited availability of monospecific alloantibodies capable of recognizing HLA epitopes [120]. Filling this gap, knowledge of epitope-paratope interaction using protein structure crystallographic models allowed for *in silico* determination of HLA epitopes [121].

The total epitope surface interfacing an antibody paratope is around 15-22 residues, with a contact area of 650-900 Å radius, but the specificity of an epitope is associated with a smaller region of 3-3.5 Å radius. Referred as 'functional' HLA epitopes, this epitope area mainly comprises of polymorphic amino acids interfacing the centrally located complementary determining region (CDR) H3 loop, which is the loop mostly associated with antibody specificity (Figure 1.3, Chapter 1). Exposed mismatched polymorphisms in the HLA molecule surface within the 'functional' epitope area are seen to be responsible for eliciting an antibody response (Figure 3.1). This set of polymorphic amino acids (or 'eplets') consist of the functional units defining an HLA epitope to be compared in HLA epitope matching.

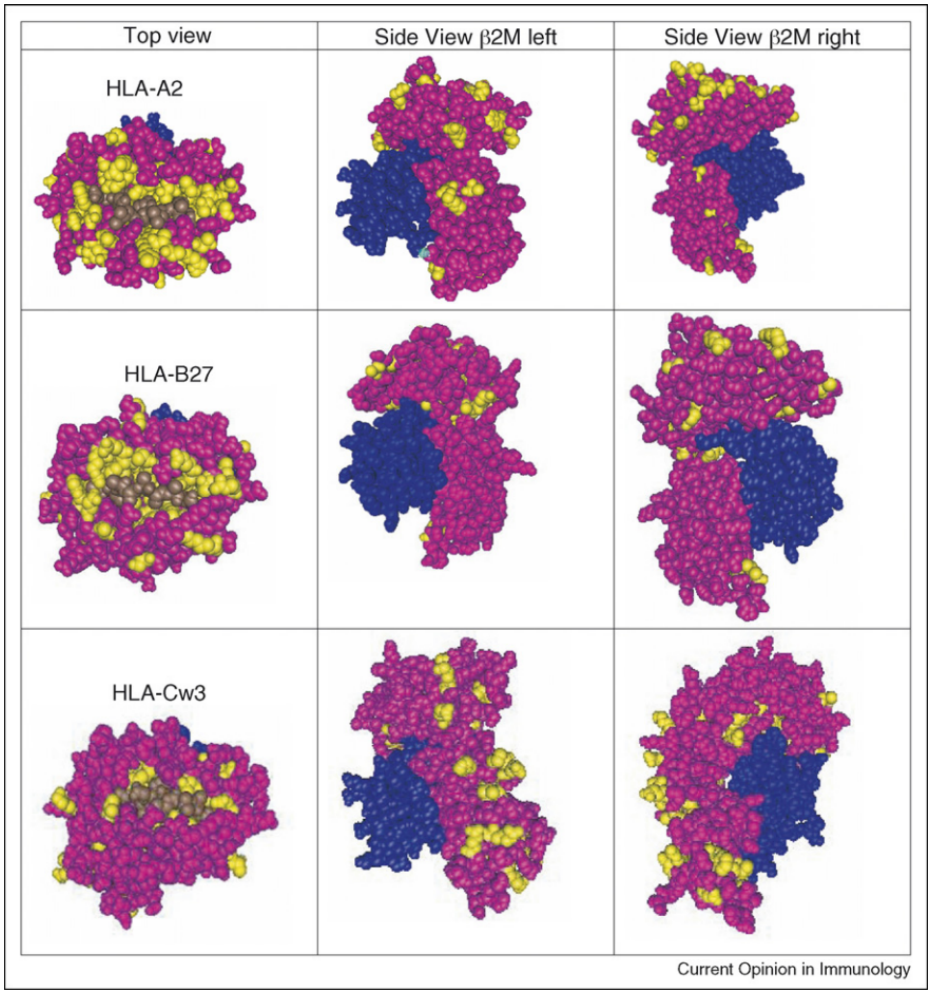


Figure 3.1: Polymorphic residues on HLA-A, -B and -C molecules. On the α-heavy chain, residues in yellow are polymorphic and residues in pink are conserved. In blue, the β2-microglobulin that is responsible for anchoring the HLA molecule in the membrane [122]

Several HLA epitopes are shared among different HLA alleles and across different HLA loci, a reflex of evolutionary mechanisms such as convergent evolution and gene recombination acting on the HLA gene family [36]. This characteristic is the main reason why some individuals become 'highly sensitized' if they are immunized against an epitope shared by many HLA alleles appearing at a high frequency in populations. Nevertheless, the extent of shared epitopes between HLA alleles and loci raises the possibility of HLA epitope compatibility between mismatched HLA types (Figure 3.2). Furthermore, HLA epitope matching enables a quantitative analysis based on the number of mismatched epitopes. Different HLA type comparisons having the same number of allele mismatches can show different numbers of HLA epitope mismatches (also referred as 'epitope load'). Since a higher number of HLA epitope mismatches have been correlated with an increased risk of rejection, [122,123] information on the number of epitope mismatches could help patients who are unable to find a 'match' to minimize their risk of rejection.

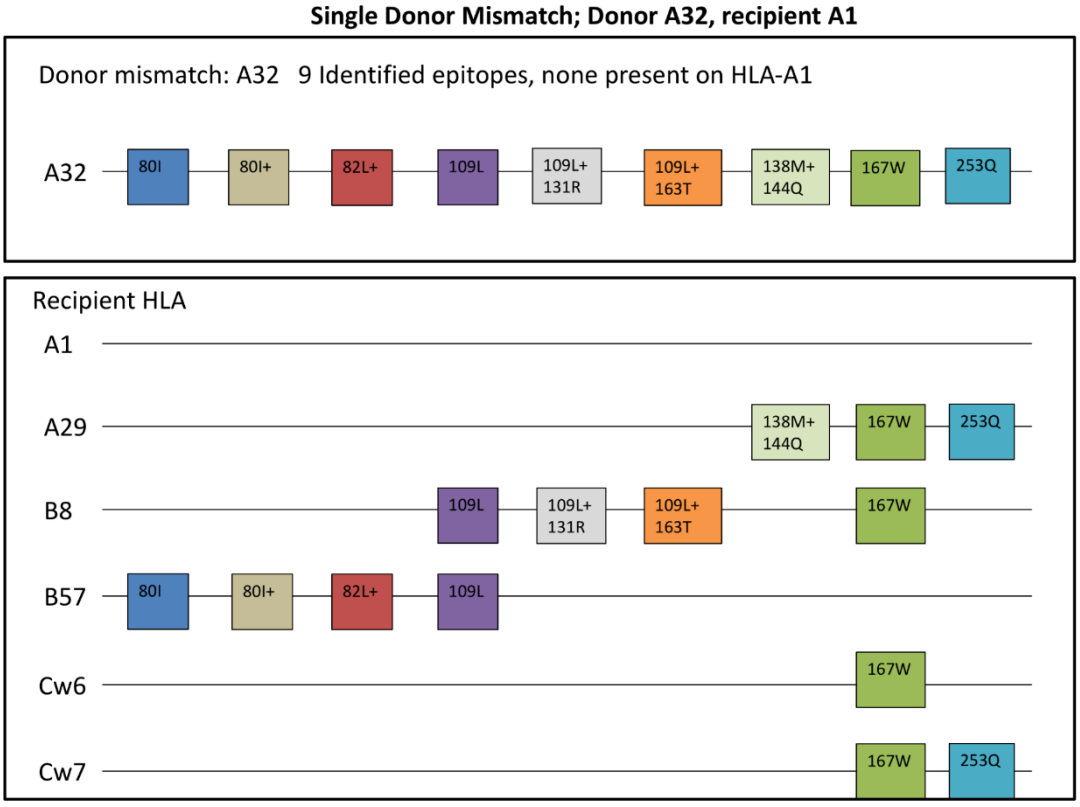
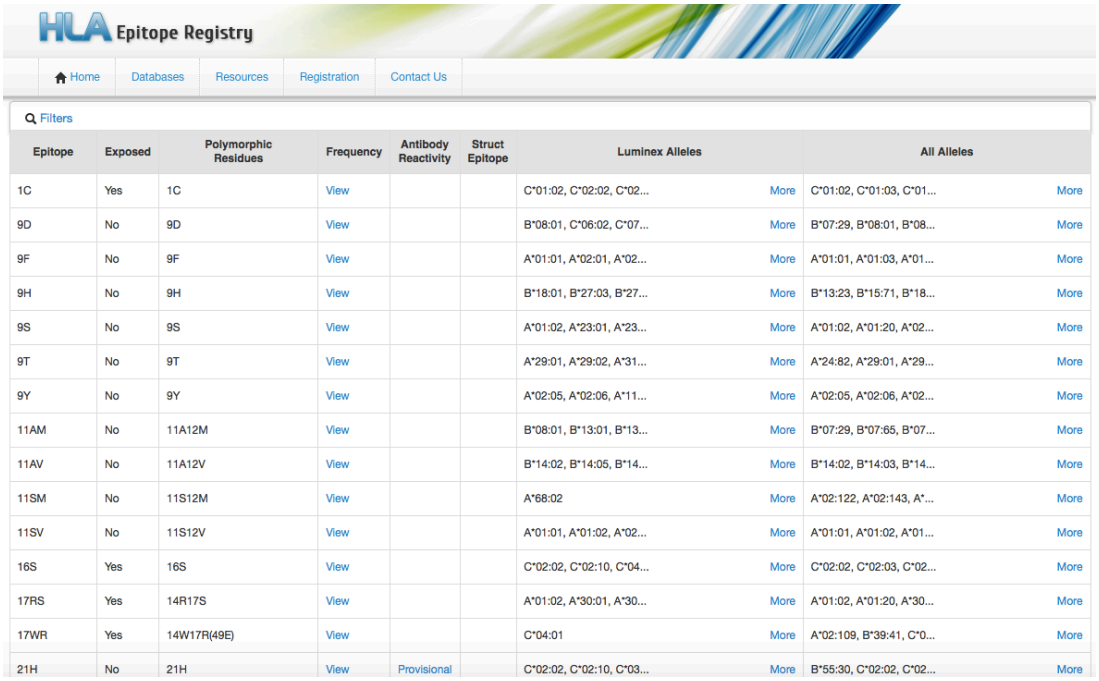


Figure 3.2: Single HLA-A antigen mismatch representing an acceptable mismatch from an HLA epitope perspective (Donor: A*32 / Patient: A*01). Despite nine epitopes in A*32 that are not present in A*01, they are configure as shared self-epitopes present in other HLA antigens from patient.

A list of HLA epitope definitions is maintained by the HLA Epitope Registry (http://www.epregistry.com.br/) [120] (Figure 3.3), with a nomenclature system based on positions and abbreviated names of polymorphic residues composing the 'eplets' of each HLA epitope. Some of the listed HLA epitopes have antibody reactivity evidence ('provisional' or 'confirmed'), while others have only been predicted *in silico*. The database also provides the HLA alleles mapped to each epitope, including a selection for Luminex® SAB alleles only. Using the aforementioned HLA epitope definitions, an epitope matching algorithm called HLA Matchmaker has been developed [124] to determine epitope mismatches in HLA types from patients and donors, and also to identify epitopes belonging to reactive HLA alleles in SAB assays possibly being recognized by alloantibodies.

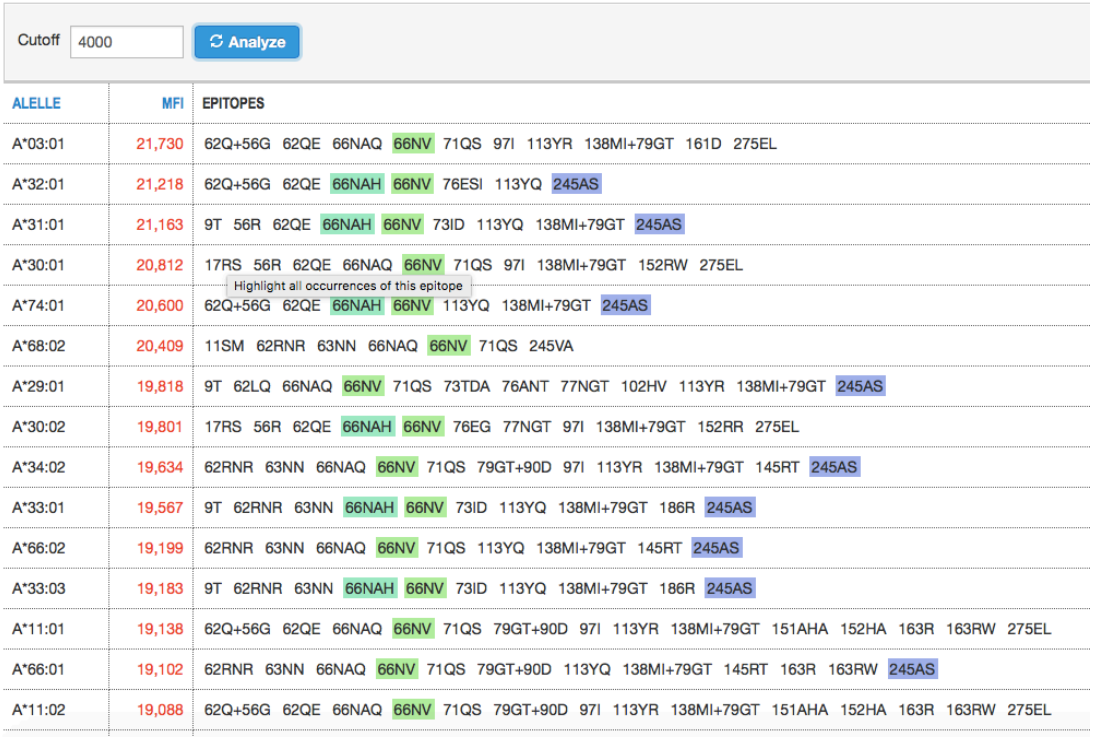


Epitope	Exposed	Polymorphic Residues	Frequency	Antibody Reactivity	Struct Epitope	Luminex Alleles		All Alleles	
1C	Yes	1C	View			C*01:02, C*02:02, C*02...	More	C*01:02, C*01:03, C*01...	More
9D	No	9D	View			B*08:01, C*06:02, C*07...	More	B*07:29, B*08:01, B*08...	More
9F	No	9F	View			A*01:01, A*02:01, A*02...	More	A*01:01, A*01:03, A*01...	More
9H	No	9H	View			B*18:01, B*27:03, B*27...	More	B*13:23, B*15:71, B*18...	More
9S	No	9S	View			A*01:02, A*23:01, A*23...	More	A*01:02, A*01:20, A*02...	More
9T	No	9T	View			A*29:01, A*29:02, A*31...	More	A*24:82, A*29:01, A*29...	More
9Y	No	9Y	View			A*02:05, A*02:06, A*11...	More	A*02:05, A*02:06, A*02...	More
11AM	No	11A12M	View			B*08:01, B*13:01, B*13...	More	B*07:29, B*07:65, B*07...	More
11AV	No	11A12V	View			B*14:02, B*14:05, B*14...	More	B*14:02, B*14:03, B*14...	More
11SM	No	11S12M	View			A*68:02	More	A*02:122, A*02:143, A*...	More
11SV	No	11S12V	View			A*01:01, A*01:02, A*02...	More	A*01:01, A*01:02, A*01...	More
16S	Yes	16S	View			C*02:02, C*02:10, C*04...	More	C*02:02, C*02:03, C*02...	More
17RS	Yes	14R17S	View			A*01:02, A*30:01, A*30...	More	A*01:02, A*01:20, A*30...	More
17WR	Yes	14W17R(49E)	View			C*04:01	More	A*02:109, B*39:41, C*0...	More
21H	No	21H	View	Provisional		C*02:02, C*02:10, C*03...	More	B*55:30, C*02:02, C*02...	More

Figure 3.3: Screenshot of the HLA Epitope Registry database, showing a subset of class I epitopes and HLA alleles where they are present, subdivided in a category containing only alleles present in Luminex® SAB and another category including alleles absent in SAB assays.

Originally, HLA Matchmaker was distributed as a spreadsheet file, but due to the difficulty to use the tool in a clinical environment because of its time-consuming analysis of alloantibodies, EpVix was conceived as a web-based version of HLA MatchMaker with a more user-friendly interface [125] (Figure 3.4). For alloantibody analysis, EpVix takes

the patient's HLA type and SAB assay result as input, as well as the chosen MFI cut-off value. All epitopes present in the patient's HLA type are flagged as self-epitopes. Reactivity to epitopes present in HLA alleles below the cut-off value is considered negative. Finally, the remaining epitopes present only in HLA alleles above the cut-off values are considered potential epitopes being targeted by alloantibodies, as in practice it is possible that only some of the remaining epitopes may be alloantibody targets. For example, an output from EpVix may show a common epitope related to all HLA alleles above the cut-off, suggesting the possibility that only that epitope has caused the patient immunization against all reactive HLA types in the SAB assay. In these situations, knowledge of the HLA types causing the immunization event helps verifying actual reactive epitopes.



ALELLE	MFI	EPITOPES
A*03:01	21,730	62Q+56G 62QE 66NAQ 66NV 71QS 97I 113YR 138MI+79GT 161D 275EL
A*32:01	21,218	62Q+56G 62QE 66NAH 66NV 76ESI 113YQ 245AS
A*31:01	21,163	9T 56R 62QE 66NAH 66NV 73ID 113YQ 138MI+79GT 245AS
A*30:01	20,812	17RS 56R 62QE 66NAQ 66NV 71QS 97I 138MI+79GT 152RW 275EL
A*74:01	20,600	62Q+56G 62QE 66NAH 66NV 113YQ 138MI+79GT 245AS
A*68:02	20,409	11SM 62RNR 63NN 66NAQ 66NV 71QS 245VA
A*29:01	19,818	9T 62LQ 66NAQ 66NV 71QS 73TDA 76ANT 77NGT 102HV 113YR 138MI+79GT 245AS
A*30:02	19,801	17RS 56R 62QE 66NAH 66NV 76EG 77NGT 97I 138MI+79GT 152RR 275EL
A*34:02	19,634	62RNR 63NN 66NAQ 66NV 71QS 79GT+90D 97I 113YR 138MI+79GT 145RT 245AS
A*33:01	19,567	9T 62RNR 63NN 66NAH 66NV 73ID 113YQ 138MI+79GT 186R 245AS
A*66:02	19,199	62RNR 63NN 66NAQ 66NV 71QS 113YQ 138MI+79GT 145RT 245AS
A*33:03	19,183	9T 62RNR 63NN 66NAH 66NV 73ID 113YQ 138MI+79GT 186R 245AS
A*11:01	19,138	62Q+56G 62QE 66NAQ 66NV 71QS 79GT+90D 97I 113YR 138MI+79GT 151AHA 152HA 163R 163RW 275EL
A*66:01	19,102	62RNR 63NN 66NAQ 66NV 71QS 79GT+90D 113YQ 138MI+79GT 145RT 163R 163RW 245AS
A*11:02	19,088	62Q+56G 62QE 66NAQ 66NV 71QS 79GT+90D 97I 113YR 138MI+79GT 151AHA 152HA 163R 163RW 275EL

Figure 3.4: Screenshot of EpVix [125] option for epitope analysis of alloreactive antibody SAB profiles generating all possible reactive epitopes, with additional options to manipulate the MFI cut-off value and interactively highlight epitopes shared by different alleles.

HLA matchmaker has been mostly applied to HLA mismatch acceptability for sensitized patients by minimizing epitope mismatches as an attempt to improve overall survival of the transplanted organ [126,127]. The Eurotransplant organisation, a program

to facilitate donor selection and cross-border exchange of donor organs, has incorporated epitope matching for highly sensitized patients [123].

With the investigation of HLA epitopes starting to be part of the routine of transplant laboratories and an added factor in donor selection programs, it is also necessary to gather knowledge of the worldwide distribution of HLA epitopes for improving matching algorithms using HLA epitope data. A population HLA epitope frequency is the percentage of the population possessing a given epitope (phenotypic frequency) that can be expressed by distinct HLA genes or alleles. Similar to how HLA allele and haplotype population frequencies are currently used in donor search algorithms [128,129], HLA epitope frequencies from populations can be used for development of epitope-based donor selection strategies.

EpFreq-DB database, described in this chapter, has been developed as a new repository within the Allele Frequency Net Database (AFND) aiming to fill this knowledge gap by storing worldwide HLA epitope frequencies originated from HLA population genotype data. A short description of EpFreq-DB has been reported in two manuscripts published in Transfusion Medicine and Hemotherapy journal (2014) [87] and in the Nucleic Acids Research journal (2015) [58] describing the latest updates in AFND (Appendix A). Furthermore, the generated HLA epitope population data was used to explore the applicability of HLA epitopes in alloantibody reactivity analysis in sensitized patients by comparing the matching likelihoods based on alloreactive HLA epitopes to likelihoods based on alloreactive HLA antigens from SAB assay results.

3.3 Methods

3.3.1 HLA Epitope Definitions

The HLA epitopes nomenclature and definitions used in EpFreq-DB are according to the first release of the HLA Epitope Registry [120]. Each epitope definition corresponds to single or multiple polymorphic residues within the molecular conformation of a 'functional' epitope, sometimes including specific self-residues that were shown to be necessary for immunogenicity of certain epitopes. A mixture of positions and amino acid letter abbreviations from their definitions are used to compose

unique nomenclatures to each HLA epitope. In the present study only epitopes belonging to HLA class I were considered, and data associated to each epitope was extracted. HLA class II epitopes have not been included since limited genotype data for HLA class II is available, being mostly HLA-DRB1 typing. For determining HLA class II epitope frequencies in the HLA-DRB locus, data regarding HLA-DRB3, -DRB4 and -DRB5 is necessary since there are shared epitopes between those genes, otherwise inaccurate estimates of epitope frequencies would be generated. The latest release of HLA Epitope Registry has updated epitope definitions partially different from the first release. The epitope matching analyses in the present study use epitope definitions from the latest version, while definitions in EpFreq-DB will be updated at a later stage.

3.3.2 HLA Population Data

HLA raw genotyping data and haplotype frequency data from AFND were used for calculating epitope frequencies, but only HLA raw genotyping data was used to generate HLA epitope profiles for matching analysis. All datasets have alleles at high resolution, i.e. with an HLA nomenclature of at least two fields (e.g. A*01:01). Although AFND has a vast amount of HLA allele frequency data, only haplotype frequency data was considered suitable for calculating epitope frequencies. Estimation of HLA epitope frequencies from allele frequencies was attempted based on probability theory, but results differed from estimations made using HLA genotypes. Frequencies from epitopes present simultaneously in different loci are influenced by population linkage disequilibrium in the HLA loci, and estimating their frequencies from allele frequencies with no LD data led to considerable variation between observed and expected frequencies (data not shown). In total, 30 populations have been included in EpFreq-DB, 16 consisting of raw genotyping data and 14 consisting of haplotype frequencies. Regarding their geographic distribution, 20% belong to Africa, 23% from Asia, 33% from Europe, 20% are from North America and 3% are from Oceania.

3.3.3 Epitope Frequency Calculation

HLA epitope frequencies were obtained from HLA raw genotyping data and HLA haplotype frequencies. Frequency calculation from HLA raw genotyping data was performed according to following steps. First, HLA epitope individual profiles were

generated by listing distinct epitopes mapping to alleles in individual HLA genotypes. Second, the frequency of each epitope mapped in a population was obtained by direct counting. Since HLA genotyping data is generally not publicly available, it is also desirable being able to use HLA haplotype frequency data for obtaining epitope frequencies. For this purpose, the logic from Hardy-Weinberg proportions (HWP) equation – a mathematical concept commonly used in population genetics studies to estimate expected genotype frequencies in a population [130] – was applied to HLA haplotype frequencies (Figures 3.5 and 3.6).

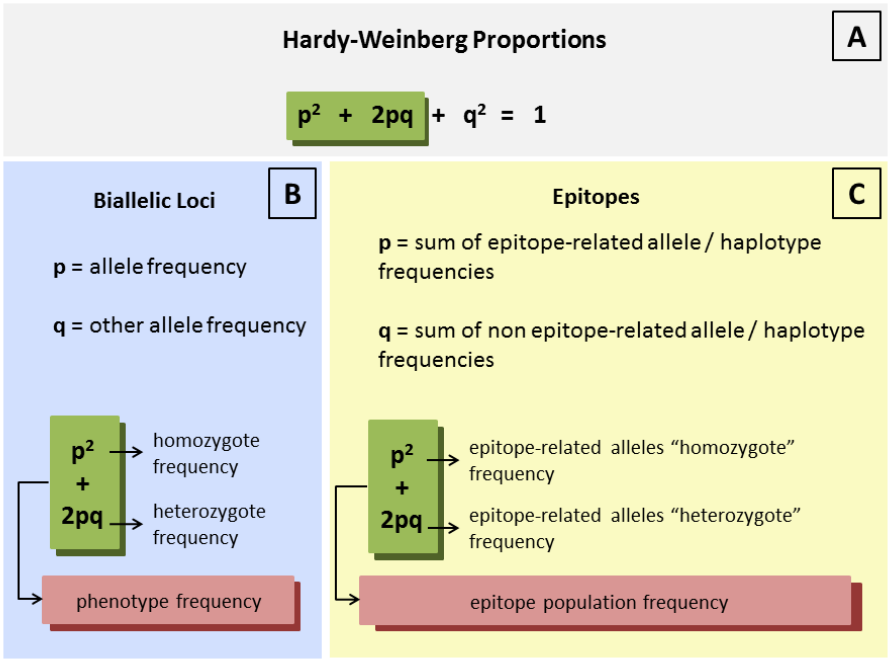


Figure 3.5: Comparison of the variables used in the application of HWP for biallelic loci or HLA epitopes. (A) The HWP equation and highlights the components p^2 and $2pq$. (B) A detailed description of the variables originated from HWP equation in a usual biallelic loci scenario. (C) Application of HWP variables in the epitope scenario, where HLA allele variants are divided in two groups: a "homozygote" frequency group corresponding to the percentage of individuals containing epitope-related alleles / haplotypes in one locus for both chromosomes, and a "heterozygote" frequency group corresponding to the percentage of individuals containing epitope-related alleles / haplotype in one locus and in only one chromosome.

Methods for Estimating Epitope Population Frequency

ENTRY DATA

England North West Population
- Typed for HLA-A, -B and –C
- N = 298

Epitope 65GK (Single Locus)
- Alleles: A*02129, A*0246, A*0248, A*0330, **A*2301**, A*2318, A*2319, A*2320, A*2321, **A*2402**, **A*2403**, A*2404, … (HLA-A only)

Epitope 66K (Multi Loci)
- Alleles: **A*0201**, A*0202, A*0203, A*0204, A*0205, A*0210, … , B*0713, B*4601, B*4602, B*4603, B*4604, B*4612, … , **C*0102**, C*0103, C*0104, C*0105, C*0106, C*0110, … , C*0202, C*0203, C*0210, … ,C*0302, **C*0303**, **C*0304**, C*0305, C*0306, C*0438, C*0440, **C*0501**, C*0620, C*0622, C*0623, **C*0702**, C*0703, C*0704, C*0767, C*0768, C*0772, C*0801, **C*0802**, C*0803, C*0822, C*0823, C*1202, **C*1203**, C*1204, C*1205, C*1206, C*1207, C*1209, C*1219, C*1220, C*1221, **C*1402**, C*1403, C*1404, C*1801, C*1802, C*1803,…

Raw Data: Direct Count of Individuals with Epitope Alleles

Sample	HLA Genotype						Epitopes	
	A		B		C		65GK	66K
Subject 1	0201	6801	4001	4402	0304	0501	Absent	Present
Subject 2	2402	6801	1501	2705	0102	0303	Present	Present
Subject 3	0101	0301	0702	0801	0701	0702	Absent	Present
Subject 4	0301	2601	0702	3801	0702	1203	Absent	Present
Subject N	Same process for each individual…							

Allele Frequencies and Haplotype Frequencies
Hardy-Weinberg Proportions
Note: Haplotype frequencies were estimated using PyPop software

Step 1: Sum of frequencies of epitope-related alleles

65GK		66K			
Allele	Allele Frequency	A	B	C	Haplotype Frequency
A*2301	0.0185	A*0101	B*0702	**C*0702**	0.0117
A*2402	0.0688	A*0101	B*0801	**C*0701**	0.1167
A*2403	0.0017	A*0101	B*1402	**C*0802**	0.0034
SUM	**0.0889**	A*0101	B*1501	**C*0102**	0.0015
		A*0101	B*1501	**C*0303**	0.0018
For the ENW population, the only alleles related to 65GK are A*2301, A*2402 and A*2403.		A*0101	B*2705	**C*0102**	0.0018
		…			
		Almost all haplotypes have 66K		**SUM**	**0.9757**

Step 2: Application Hardy-Weinberg Proportions

Epitope	Sum of Frequencies*	$p^2 + 2pq$
65GK	0.0889	$0.0889^2 + 2*0.0889*(1-0.0889)$
66K	0.9757	$0.9757^2 + 2*0.9757*(1-0.9757)$

RESULTS – Population Frequency of Epitopes

Direct Count of Individuals with Epitope Alleles:
- 65GK = 0.1779 66K = 1.000

Estimation from HLA allele frequencies
- 65GK = 0.1700 66K = 0.9994

Figure 3.6: Example of all the steps involved in the calculation of population frequency of 65GK and 66K epitopes in England North West population, using both types of entry data (HLA raw data and allele/haplotype frequencies). In the raw data box, the alleles having a 66K epitope are outlined in red and the alleles having a 65GK epitope are filled in yellow.

For any given biallelic locus, HWP equation can be applied to estimate its expected genotype frequencies (p^2, $2pq$ and q^2 frequencies) using population frequencies from both

alleles (p and q) (Figure 3.5). Applying the probability theory logic behind HWP, all HLA haplotypes in a population were classified in two categories according to the presence and absence of a given epitope. Similar to biallelic loci, the frequency of both categories represented p and q variables (p = epitope is present in the haplotype, q = epitope is absent in the haplotype). Lastly, the epitope frequency can be extracted from HWP equation as the percentage of individuals in a population possessing the epitope in any of their haplotypes (Epitope frequency = $p^2 + 2pq$). An example showing the application of both methods is summarized in Figure 3.6. Perl scripts were developed to perform the HLA epitope frequency calculations. Additional hierarchical clustering analysis of epitope frequencies distribution was performed using R Statistical Programming Language [89].

3.3.4 EpFreq-DB Implementation

The back-end of EpFreq-DB database was developed using a Microsoft SQL Server relational database schema. The organization and relationship of core information stored in the database is described by the entity-relationship diagram presented in Figure 3.7. HLA epitopes and their respective characteristics were defined as the following relational entities in the database schema: 'Epitope' containing epitope names according to HLA Epitope Registry database, 'Polypeptide' containing structural mapping of epitopes (position and amino acid), 'Allele' mapping their polymorphisms, 'Locus' mapping their genetic locations, and 'Population' and 'Frequency' containing geographical information of populations and their epitope frequencies.

Users can connect to the database using the most common web browsers. A web interface for EpFreq-DB has been created to allow users to query the database, being linked to the Allele Frequency Net Database and following a similar web design. For that purpose, interactive web pages for querying data were developed using the Active Server Pages (ASP) scripting environment and JavaScript language. The graphical display was designed using HyperText Markup Language (HTML) and Cascading Style Sheets (CSS), ensuring that the page will be viewable in most used web browsers.

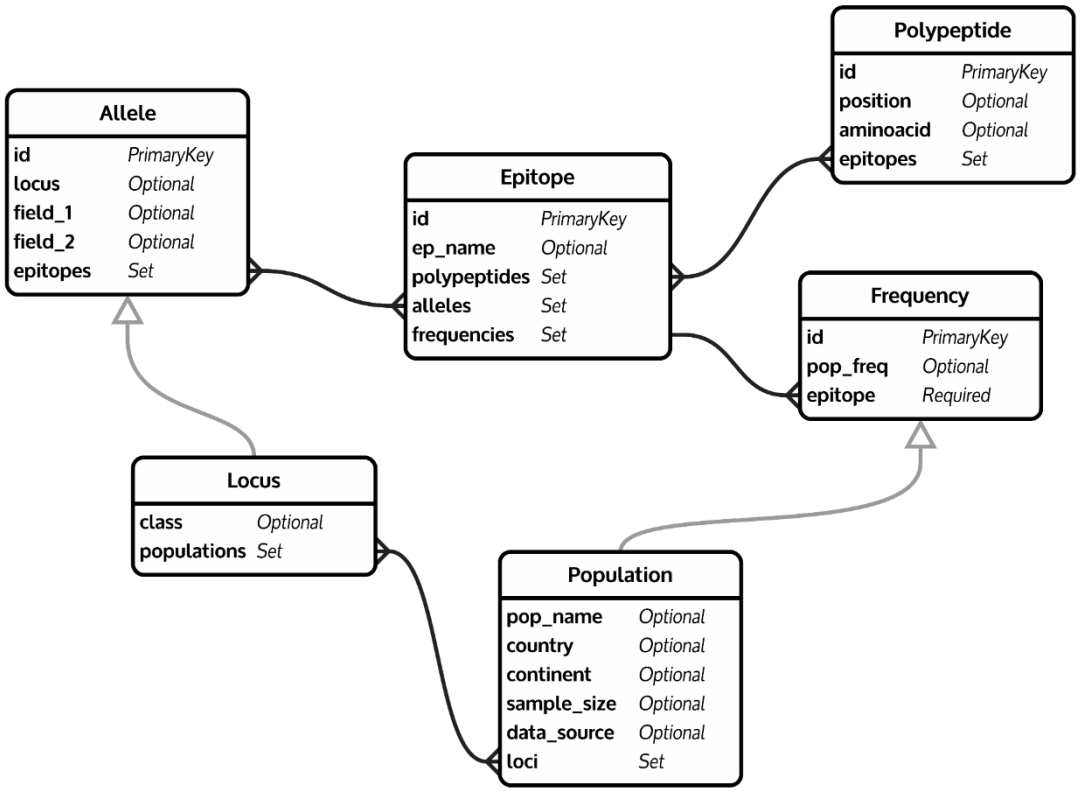


Figure 3.7: Entity-relationship diagram of EpFreq-DB schema.

3.3.5 HLA epitope alloreactivity analysis

An analysis comparing HLA epitope matching to traditional HLA allele matching regarding sensitization by anti-HLA DSA was performed using scripts developed in Python programming language. In summary, alloantibody profiles from SAB assay results from three patients were analysed in EpVix for determination of HLA epitopes possibly recognized by detected alloantibodies. Then, the percentage of individuals in each population not possessing any of those epitopes was then compared to the percentage of individuals not having the HLA types considered reactive in SAB assay results. The details involved in each step of this analysis are described hereafter.

HLA epitope profiles from populations

Available HLA raw genotype data from AFND populations was converted into HLA epitope profiles, i.e. a list containing all distinct HLA epitopes belonging to each individual's HLA genotype from each population was generated. Only HLA class I types

(HLA-A, -B and -C) were converted. Table 3.1 shows detailed information from populations, including geographical information, HLA genes typed and sample size. Several populations do not have HLA typing data for all class II genes, such as those associated to renal transplant requiring only HLA-A, -B and -DRB1. Since some epitopes can be simultaneously originated from more than one locus, selection of HLA epitopes composing profiles of each population was made in accordance to HLA genes typed. If an epitope is expressed by alleles part of a HLA gene not typed in a population, the epitope is excluded from the conversion, even if it can be present in other genes with typing information available. As an example, if a population does not have HLA-C typing information, epitopes that can be expressed by HLA-C alleles are excluded, even if they are unique to HLA-C alleles, or are also present in HLA-A and / or HLA-B alleles. This same logic has been also applied when obtaining epitope frequencies. This step prevents inaccurate estimates regarding the likelihood of finding an individual with a given epitope in the population due to its undetectable presence in non-typed HLA loci.

Table 3.1: Populations having HLA raw genotyping data used to generate epitope profiles.

Population	Continent	Typed Loci	N
Ireland Northern	Europe	A, B, C	1000
England North West	Europe	A, B, C	298
Netherlands UMCU	Europe	A, B, C	64
Malaysia Peninsular Malay	South-East Asia	A, B, C	951
Portugal Azores Terceira Island	Europe	A, B, C	130
Malaysia Peninsular Chinese	South-East Asia	A, B, C	194
New Zealand Polynesians with Full Ancestry	Oceania	A, B, C	21
New Zealand Polynesians with Admixed History	Oceania	A, B, C	27
New Zealand Maori with Full Ancestry	Oceania	A, B, C	46
New Zealand Maori with Admixed History	Oceania	A, B, C	105
Malaysia Kedah Kensiu	South-East Asia	A, B	21
Hong Kong Chinese BMDR	South-East Asia	A, B, C	7795
Hong Kong Chinese cord blood registry	South-East Asia	A, B	3892
Albania DKMS	Europe	A, B, C	146
Austria DKMS	Europe	A, B, C	854
Belarus DKMS	Europe	A, B, C	42
Belgium DKMS	Europe	A, B, C	199
Bosnia DKMS	Europe	A, B, C	399
Bulgaria DKMS	Europe	A, B, C	149
Croatia DKMS	Europe	A, B, C	801

Czech DKMS	Europe	A, B, C	310
Denmark DKMS	Europe	A, B, C	69
Estonia DKMS	Europe	A, B, C	21
Faroe DKMS	Europe	A, B, C	64
Finland DKMS	Europe	A, B, C	64
France DKMS	Europe	A, B, C	530
Greece DKMS	Europe	A, B, C	749
Hungary DKMS	Europe	A, B, C	256
Ireland DKMS	Europe	A, B, C	57
Italy DKMS	Europe	A, B, C	2224
Latvia DKMS	Europe	A, B, C	33
Lithuania DKMS	Europe	A, B, C	66
Luxemburg DKMS	Europe	A, B, C	118
Macedonia DKMS	Europe	A, B, C	87
Montenegro DKMS	Europe	A, B, C	45
Nederlands DKMS	Europe	A, B, C	583
Norway DKMS	Europe	A, B, C	24
Poland DKMS	Europe	A, B, C	3250
Portugal DKMS	Europe	A, B, C	429
Romania DKMS	Europe	A, B, C	552
Russia DKMS	Europe	A, B, C	2227
Serbia DKMS	Europe	A, B, C	283
Slovakia DKMS	Europe	A, B, C	70
Slovenia DKMS	Europe	A, B, C	130
Spain DKMS	Europe	A, B, C	627
Sweden DKMS	Europe	A, B, C	62
Swizterland DKMS	Europe	A, B, C	266
Ukraine DKMS	Europe	A, B, C	269
UK DKMS	Europe	A, B, C	419

Identification of HLA epitopes specific to anti-HLA alloantibodies

HLA alloantibody profiles pre-loaded as sample data in EpVix tool from three patients were used as the input dataset matched against populations. The profiles were in Luminex® SAB assay format and showed variable levels of sensitization (Table 3.2). Establishment of reactive HLA alleles and epitopes was based on two MFI cut-off values: 2000 and 4000. The lower MFI cut-off value is currently used by the Transplant Immunology laboratory in the Royal Liverpool University Hospital (Middleton, D.,

personal communication), being also a common choice among other laboratories. A higher MFI cut-off value was also applied to evaluate its effect, since the selection of 2000 as "standard" cut-off is based on relatively arbitrary criteria. Reactive HLA alleles from all three patients were converted to epitopes using the HLA Matchmaker algorithm on EpVix, generating a total of two epitope datasets per patient (one dataset for each MFI cut-off value).

Minimum HLA epitope sets

In HLA epitope sets derived from EpVix, there is the possibility that not all listed epitopes are causing alloreactivity. For example, in case an epitope is common to all HLA alleles considered reactive, there is a chance that the patient's anti-HLA alloantibodies are only specific to that epitope. To evaluate the effect of selecting only a subset of epitopes capable of explaining all reactive HLA types, epitope lists generated by EpVix were reduced to the minimum epitope set common to the maximum number of reactive HLA types (Figure 3.8). Both minimum and full epitope sets were included in the analysis.

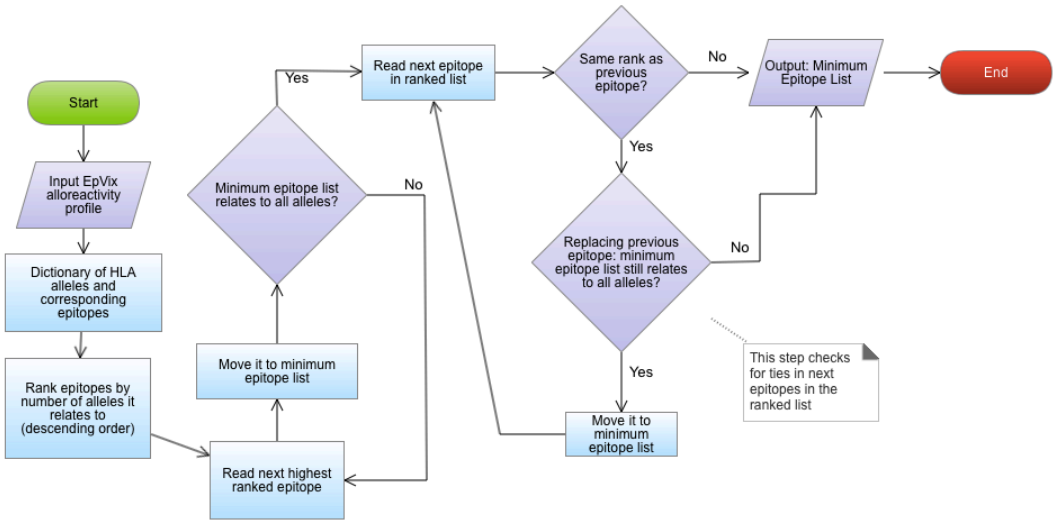


Figure 3.8: Algorithm to determine minimum HLA epitope set capable of explaining reactive alleles in Luminex® SAB assays. HLA epitopes were ranked in descending order according to the number of alleles they were related to in EpVix analyses. Iterations through the ranked list selected epitopes in the list order until all alleles have corresponding epitopes in the selected list. Additional steps were added to check for ties among epitopes with equal rank and similar ability to explain reactivity.

HLA allele and epitope 'negative' matching in populations

A data pipeline was developed in Python programming language using alloantibody reactivity profiles (for both epitopes and alleles) as input to be matched against HLA population data (epitope profiles and genotypes). Input epitope sets were also adapted to the matched populations according to their typed HLA genes. If an epitope is present in a HLA gene that has not been typed for a population, the epitope is excluded from the input set being matched against that specific population. Output values from the data pipeline correspond to the percentage of the population possessing none of the items in the input set (HLA alleles or epitopes), i.e. individuals not displaying putative units causing alloreactivity in patients ('negative matching').

3.4 Results and Discussion

3.4.1 Validation of HLA epitope frequencies estimated from HLA haplotype frequencies

To validate the calculation of HLA epitope frequencies from haplotype frequencies, the method was applied on haplotype frequencies generated from HLA raw genotyping data using PyPop software [33]. A comparison of HLA epitope frequencies obtained from generated haplotype frequencies against observed epitope frequencies calculated directly from raw genotype data were performed for populations in EpFreq-DB with available raw data. Figure 3.9 compares epitope frequencies calculated from raw genotyping data to obtained haplotype frequency data for 30 populations with sample size greater than 100, showing similar frequencies for both groups (for detailed comparison for individual populations, see Appendix C).

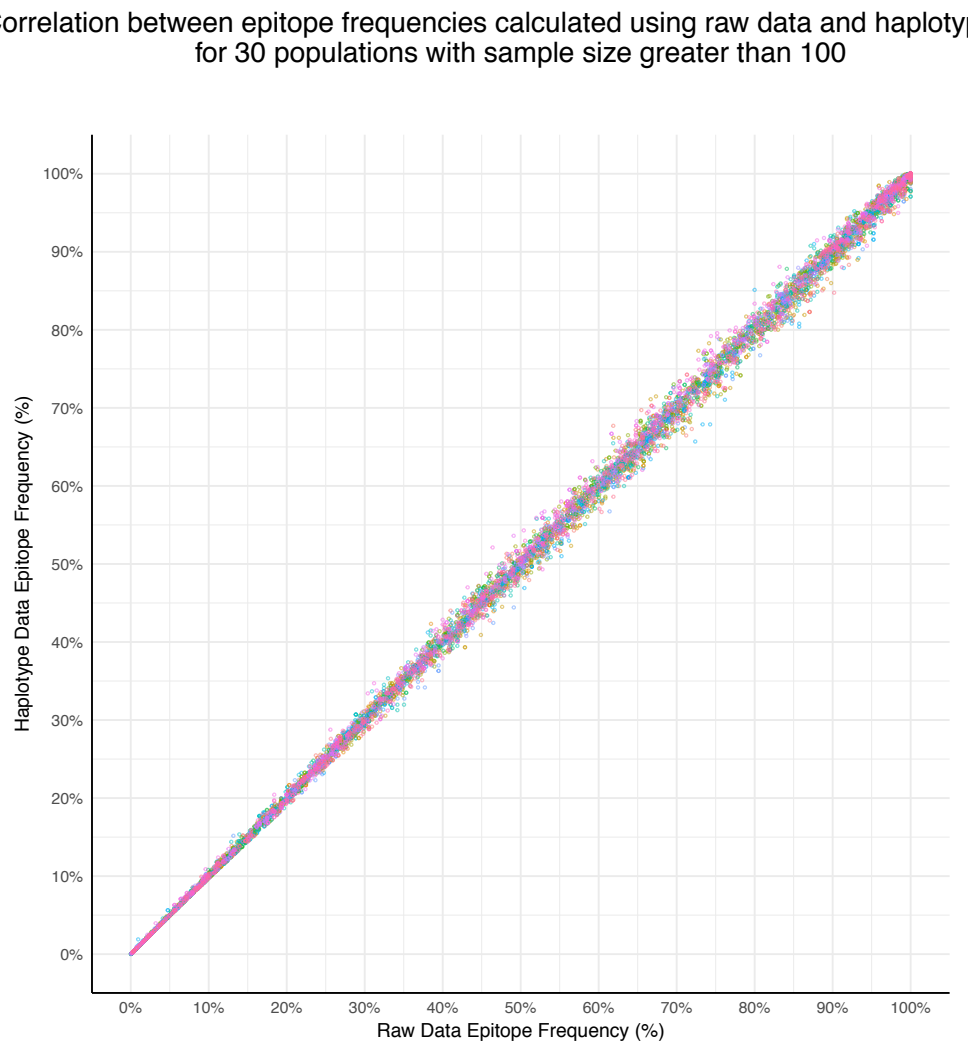


Figure 3.9: Correlation between HLA epitope frequencies calculated from HLA raw data and HLA haplotype frequency data for populations with sample size greater than 100. Each point corresponds to the population frequencies of a distinct HLA epitope calculated using HLA raw genotype data (x-axis) and HLA haplotype frequency data (y-axis), where different colours represent different populations. The observed distribution of the points shows that there is little variation between HLA epitope frequencies calculated using either data sources, for populations with sample size of at least 100 (for detailed comparison for individual populations, see Appendix C).

1.1.1 Website organization

The EpFreq-DB database is part of Allele Frequencies Net Database (AFND), and can be accessed through the AFND homepage (http://www.allelefrequencies.net/) using the menu "HLA Epitopes" and the submenu "HLA Epitopes - ABC" or via a direct URL access at http://www.allelefrequencies.net/hlaepitopes/, both options directing to

EpFreq-DB query page (Figure 3.10). The website interface allows the user to retrieve population HLA epitope frequencies applying a collection of data filters through different view modes. Retrieved data in the query page can be filtered by: i) one or more specific populations through a drop-down list to select one specific population or a textbox to select more than one population, ii) one or more specific HLA epitopes by also a drop-down list and textbox for selection, iii) demographic information (country, continent, ethnicity and sample size), iv) HLA loci from where HLA epitopes are expressed (including multiple loci combinations), v) amino acid position range in the polypeptide sequence, and vi) HLA alleles or a complete HLA genotype, retrieving data from all epitopes present in HLA alleles. It is also possible to sort results by populations or epitope names in alphabetical order or in descending order from the highest frequencies (Figure 3.10).

The default display of any retrieved data consists of a list showing population epitope frequencies per row, along with additional demographic and molecular data related to the population and HLA epitope, respectively. Results can be display in alternative views: i) a printer-friendly view similar to the default view, but with less formatting elements, ii) a grid view, showing results in a table format with populations as columns and epitopes as rows, and iii) a heatmap view, which is the grid view with cells displaying colours with their intensity varying according to the frequency value (Figure 3.11). Utilization of the grid or heatmap view in combination with filtering HLA epitopes present in specific alleles or genotype are particularly interesting for identification of populations where a set of epitopes can be found at the desired frequency: higher frequency when searching for individuals with similar epitope profiles of a patient with known HLA genotype, and lower frequencies when searching for individuals not having reactive epitopes to alloantibodies.

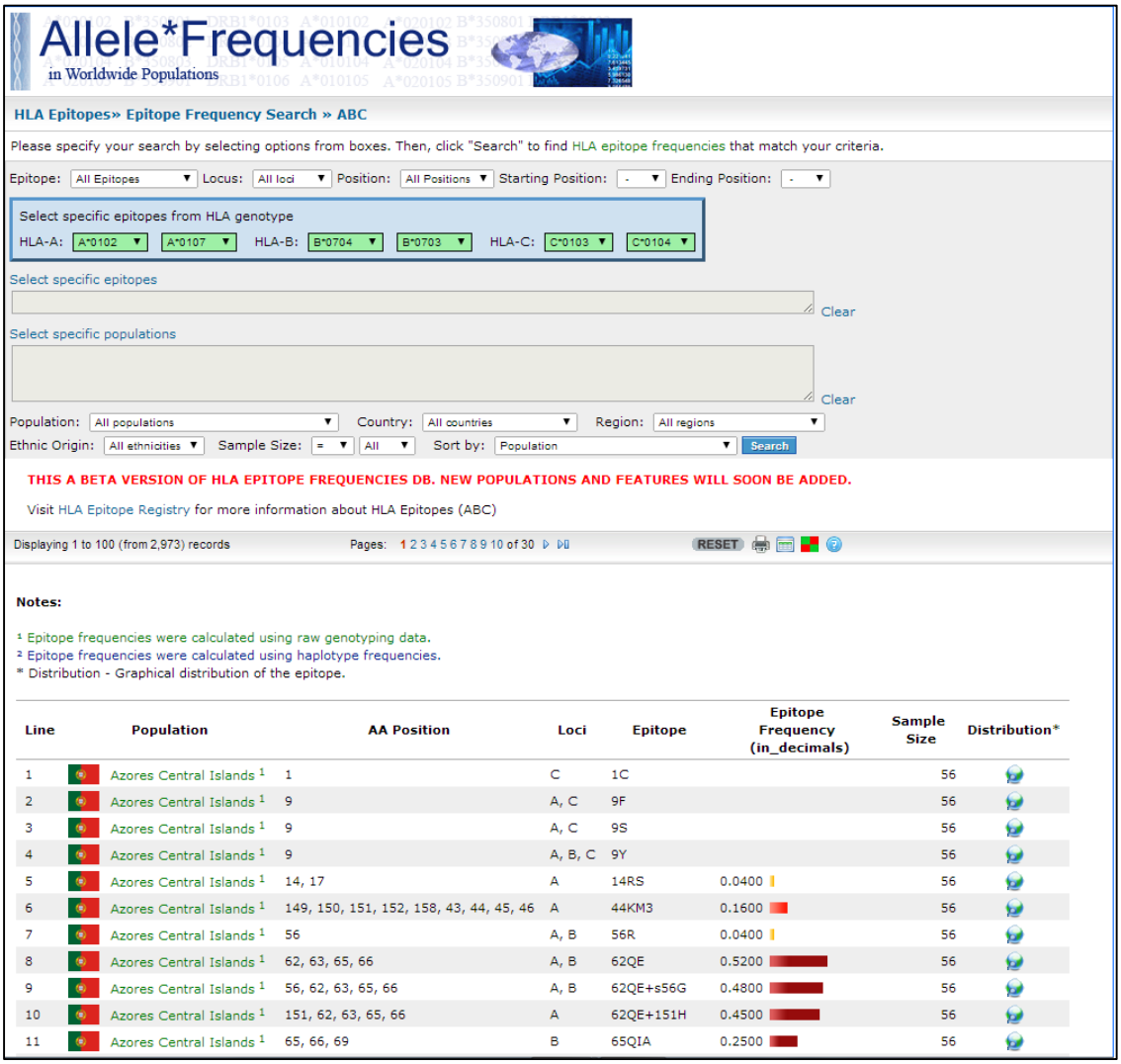
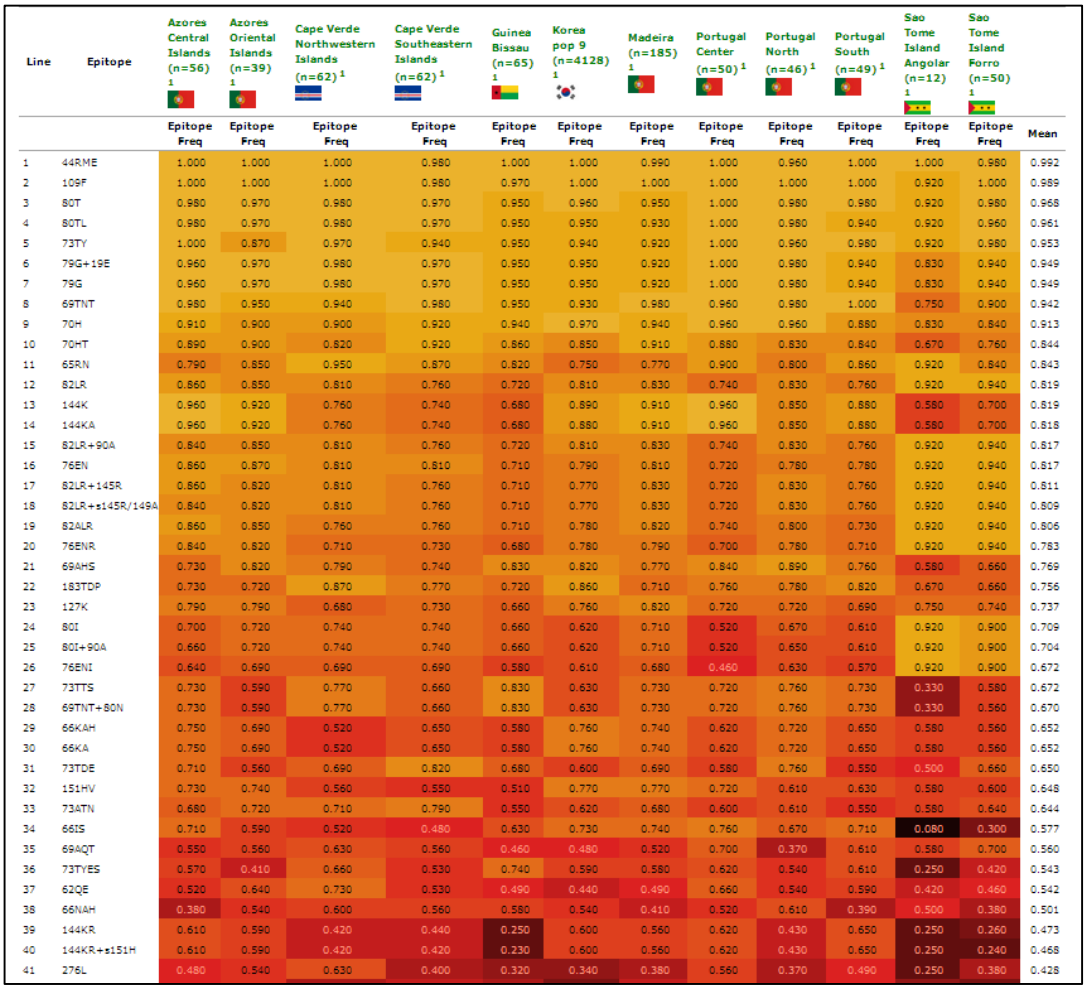


Figure 3.10: EpFreq-DB query page. Several filters for performing a specific search are available at the top of the page. A search can also be performed using no filters selected. In this example, 'HLA genotype' filters were selected (highlighted in green). The resulting epitope frequencies for each population in the database will contain only epitopes present in the specified genotype.



Line	Epitope	Azores Central Islands (n=56) 1	Azores Oriental Islands (n=39) 1	Cape Verde Northwestern Islands (n=62) 1	Cape Verde Southeastern Islands (n=62) 1	Guinea Bissau (n=65) 1	Korea pop 9 (n=4128) 1	Madeira (n=185) 1	Portugal Center (n=50) 1	Portugal North (n=46) 1	Portugal South (n=49) 1	Sao Tome Island Angolar (n=12) 1	Sao Tome Island Forro (n=50) 1	
		Epitope Freq	Epitope Freq	Epitope Freq	Epitope Freq	Epitope Freq	Epitope Freq	Epitope Freq	Epitope Freq	Epitope Freq	Epitope Freq	Epitope Freq	Epitope Freq	Mean
1	44RME	1.000	1.000	1.000	0.980	1.000	1.000	0.990	1.000	0.960	1.000	1.000	0.980	0.992
2	109F	1.000	1.000	1.000	0.980	0.970	1.000	1.000	1.000	1.000	1.000	0.920	1.000	0.989
3	80T	0.980	0.970	0.980	0.970	0.950	0.960	0.950	1.000	0.980	0.980	0.920	0.980	0.968
4	80TL	0.980	0.970	0.980	0.970	0.950	0.950	0.930	1.000	0.980	0.940	0.920	0.960	0.961
5	73TY	1.000	0.870	0.970	0.940	0.950	0.940	0.920	1.000	0.960	0.980	0.920	0.980	0.953
6	79G+19E	0.960	0.970	0.980	0.970	0.950	0.950	0.920	1.000	0.980	0.940	0.830	0.940	0.949
7	79G	0.960	0.970	0.980	0.970	0.950	0.950	0.920	1.000	0.980	0.940	0.830	0.940	0.949
8	69TNT	0.980	0.950	0.940	0.980	0.950	0.930	0.980	0.960	0.980	1.000	0.750	0.900	0.942
9	70H	0.910	0.900	0.900	0.920	0.940	0.970	0.940	0.960	0.960	0.880	0.830	0.840	0.913
10	70HT	0.890	0.900	0.820	0.920	0.860	0.850	0.910	0.880	0.830	0.840	0.670	0.760	0.844
11	65RN	0.790	0.850	0.950	0.870	0.820	0.750	0.770	0.900	0.800	0.860	0.920	0.840	0.843
12	82LR	0.860	0.850	0.810	0.760	0.720	0.810	0.830	0.740	0.830	0.760	0.920	0.940	0.819
13	144K	0.960	0.920	0.760	0.740	0.680	0.890	0.910	0.960	0.850	0.880	0.580	0.700	0.819
14	144KA	0.960	0.920	0.760	0.740	0.680	0.880	0.910	0.960	0.850	0.880	0.580	0.700	0.818
15	82LR+90A	0.840	0.850	0.810	0.760	0.720	0.810	0.830	0.740	0.830	0.760	0.920	0.940	0.817
16	76EN	0.860	0.870	0.810	0.810	0.710	0.790	0.810	0.720	0.780	0.780	0.920	0.940	0.817
17	82LR+145R	0.860	0.820	0.810	0.760	0.710	0.770	0.830	0.720	0.830	0.760	0.920	0.940	0.811
18	82LR+s145R/149A	0.840	0.820	0.810	0.760	0.710	0.770	0.830	0.720	0.830	0.760	0.920	0.940	0.809
19	82ALR	0.860	0.850	0.760	0.760	0.710	0.780	0.820	0.740	0.800	0.730	0.920	0.940	0.806
20	76ENR	0.840	0.820	0.710	0.730	0.680	0.780	0.790	0.700	0.780	0.710	0.920	0.940	0.783
21	69AHS	0.730	0.820	0.790	0.740	0.830	0.820	0.770	0.840	0.890	0.760	0.580	0.660	0.769
22	183TDP	0.730	0.720	0.870	0.770	0.720	0.860	0.710	0.760	0.780	0.820	0.670	0.660	0.756
23	127K	0.790	0.790	0.680	0.730	0.660	0.760	0.820	0.720	0.720	0.690	0.750	0.740	0.737
24	80I	0.700	0.720	0.740	0.740	0.660	0.620	0.710	0.520	0.670	0.610	0.920	0.900	0.709
25	80I+90A	0.660	0.720	0.740	0.740	0.660	0.620	0.710	0.520	0.650	0.610	0.920	0.900	0.704
26	76ENI	0.640	0.690	0.690	0.690	0.580	0.610	0.680	0.460	0.630	0.570	0.920	0.900	0.672
27	73TTS	0.730	0.590	0.770	0.660	0.830	0.630	0.730	0.720	0.760	0.730	0.330	0.580	0.672
28	69TNT+80N	0.730	0.590	0.770	0.660	0.830	0.630	0.730	0.720	0.760	0.730	0.330	0.560	0.670
29	66KAH	0.750	0.690	0.520	0.650	0.580	0.760	0.740	0.620	0.720	0.650	0.580	0.560	0.652
30	66KA	0.750	0.690	0.520	0.650	0.580	0.760	0.740	0.620	0.720	0.650	0.580	0.560	0.652
31	73TDE	0.710	0.560	0.690	0.820	0.680	0.600	0.690	0.580	0.760	0.550	0.500	0.660	0.650
32	151HV	0.730	0.740	0.560	0.550	0.510	0.770	0.770	0.720	0.610	0.630	0.580	0.600	0.648
33	73ATN	0.680	0.720	0.710	0.790	0.550	0.620	0.680	0.600	0.610	0.550	0.580	0.640	0.644
34	66IS	0.710	0.590	0.520	0.480	0.630	0.730	0.740	0.760	0.670	0.710	0.080	0.300	0.577
35	69AQT	0.550	0.560	0.630	0.560	0.460	0.480	0.520	0.700	0.370	0.610	0.580	0.700	0.560
36	73TYES	0.570	0.410	0.660	0.530	0.740	0.590	0.580	0.620	0.540	0.610	0.250	0.420	0.543
37	62QE	0.520	0.640	0.730	0.530	0.490	0.440	0.490	0.660	0.540	0.590	0.420	0.460	0.542
38	66NAH	0.380	0.540	0.600	0.560	0.580	0.540	0.410	0.520	0.610	0.390	0.500	0.380	0.501
39	144KR	0.610	0.590	0.420	0.440	0.250	0.600	0.560	0.620	0.430	0.650	0.250	0.260	0.473
40	144KR+s151H	0.610	0.590	0.420	0.420	0.230	0.600	0.560	0.620	0.430	0.650	0.250	0.240	0.468
41	276L	0.480	0.540	0.630	0.400	0.320	0.340	0.380	0.560	0.370	0.490	0.250	0.380	0.428

Figure 3.11: 'Heatmap' view of EpFreq-DB query results. The lighter the colours the higher the frequencies, and vice-versa.

3.4.2 Worldwide HLA epitope frequency distribution

Although the database web interface produces a heatmap view of world HLA epitope frequencies, a similar analysis was performed with the addition of hierarchical clustering analysis regarding the similarity of populations according to epitope frequencies and grouping of epitopes with similar frequencies in all populations. Only HLA-A and HLA-B epitopes were included in this analysis, since some populations do not have HLA-C data available. Figure 3.12 shows that epitopes have different degrees of variability across populations, where some epitopes are highly frequent in all populations, while others are consistently rare. It is also possible to identify some epitopes dominant in specific geographic regions in comparison to others.

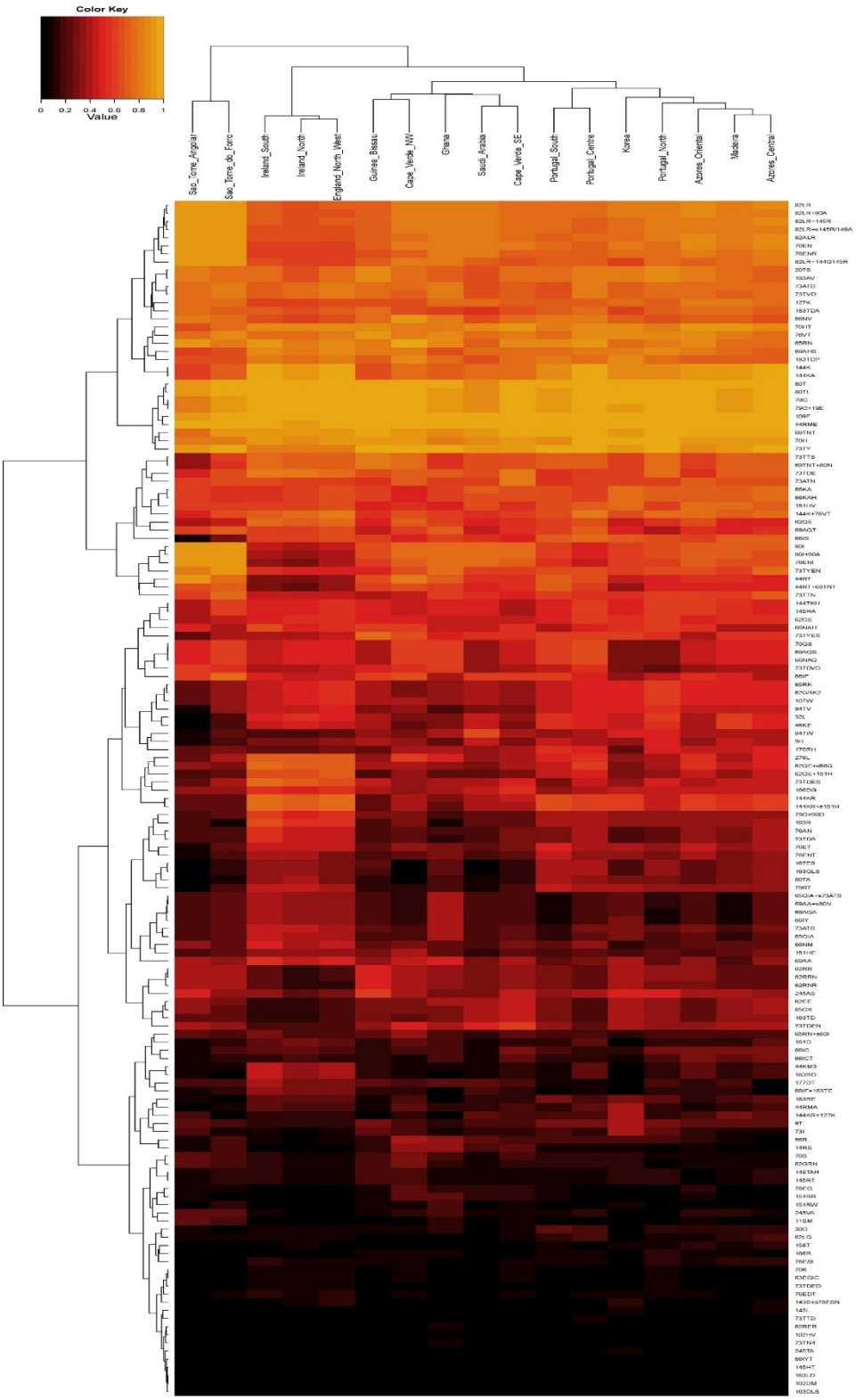


Figure 3.12: Heatmap of HLA-A and -B epitope frequencies in world populations. Lighter areas in yellow represent higher frequencies, while darker areas represent lower frequencies. Areas in red have intermediate frequencies.

Epitope frequencies in EpFreq-DB are current linked to HLA Epitope Registry database, where each HLA epitope row provides a URL link to EpFreq-DB query page specific to that epitope. With the incorporation of HLA epitope matching in transplant programs, information of their population frequencies is necessary for donor search strategies and for assessing the impact of alloreactive epitopes on patient's matching likelihood in specific populations, with the potential to improve patients' chances of finding a match in transplant programs implementing cross-border exchange of deceased donor organs.

3.4.3 HLA allele and epitope matching of alloreactivity profiles

To evaluate the impact of the use of HLA epitopes in alloreactivity profiles from sensitized patients, explorative analysis of alloreactivity profile matching comparing HLA epitopes and high resolution HLA allele matching was performed. EpVix uses the HLA MatchMaker algorithm to generates a list containing all possible alloreactive epitopes based on Luminex® SAB assay results. Alloreactivity profiles from patient samples present on EpVix with 2000 and 4000 MFI cut-off values are described in Table 3.2.

Since it may be possible that not all epitopes identified in EpVix are actual alloantibody targets, a list containing the minimum set of epitopes capable of explaining the reactive alleles from SAB assays were generated for each patient. Except for a few populations, there is no difference between the matching likelihoods obtained by both epitope sets (full and minimum epitope sets) (Figure 3.13). Differences higher than 10% between those two groups were observed only for three Asian populations (Figure 3.14). This result shows existence of non-random association between epitopes. Those selected as the minimum epitope sets are by definition the more highly frequent epitopes, serving in practice as markers highly linked to other epitopes in the full set. Following on from this result, the later analyses will only include the full epitope set to avoid influence of differences observed in outlier populations.

Table 3.2: HLA epitope analysis of three alloreactivity profiles from EpVix, using 2000 and 4000 MFI cut-off values.

Allele	MFI	HLA Epitopes (MFI cut-off 2000)	HLA Epitopes (MFI cut-off 4000)
Patient 1 (cPRA 55.86%)			
A*02:01	10368	62GE 62GK2 94TV 107W 144TKH 145KHA	62GK2 94TV 107W 144TKH 145KHA
A*02:06	9257	62GE 62GK2 94TV 107W 144TKH 145KHA	62GK2 94TV 107W 144TKH 145KHA
A*02:03	8100	62GE 62GK2 94TV 107W 144TKH 145HT	62GK2 94TV 107W 144TKH 145HT
A*68:01	7961	144TKH 145KHA 245VA	144TKH 145KHA 245VA
A*69:01	7488	94TV 107W 144TKH 145KHA	94TV 107W 144TKH 145KHA
A*68:02	6598	11SM 144TKH 145KHA 245VA	11SM 144TKH 145KHA 245VA
B*57:03	4405	62GE 62GRN 97V	None
B*57:01	3613	62GE 62GRN 97V	
A*29:01	2990	102HV	
A*24:03	2365	None	
A*30:02	2228	76EG	
B*58:01	2158	62GE 62GRN	
Patient 2 (cPRA 69.99%)			
A*03:01	21730	62QE 62QE+151H 62QE+s56G 66NAQ 66NV 71QS 97I 113YR 161D 275EL	62QE 62QE+151H 62QE+s56G 66NAQ 66NV 71QS 97I 113YR 161D 275EL
A*32:01	21218	62QE 62QE+s56G 66NAH 66NV 76ESI 113YQ 245AS	62QE 62QE+s56G 66NAH 66NV 76ESI 113YQ 245AS
A*31:01	21163	9T 56R 62QE 66NAH 66NV 73ID 113YQ 245AS	9T 56R 62QE 66NAH 66NV 73ID 113YQ 245AS
A*30:01	20812	17RS 56R 62QE 66NAQ 66NV 71QS 97I 152RW 275EL	17RS 56R 62QE 66NAQ 66NV 71QS 97I 152RW 275EL
A*74:01	20600	62QE 62QE+s56G 66NAH 66NV 113YQ 245AS	62QE 62QE+s56G 66NAH 66NV 113YQ 245AS
A*68:02	20409	11SM 62RNR 62RR 62RRN 63NN 66NAQ 66NV 71QS 245VA	11SM 62RNR 63NN 66NAQ 66NV 71QS 245VA

A*29:01	19818	9T 62LQ 66NAQ 66NV 71QS 73TDA 76ANT 77NGT 102HV 113YR 245AS	9T 62LQ 66NAQ 66NV 71QS 73TDA 76ANT 77NGT 102HV 113YR 245AS
A*30:02	19801	17RS 56R 62QE 66NAH 66NV 76EG 77NGT 97I 152RR 275EL	17RS 56R 62QE 66NAH 66NV 76EG 77NGT 97I 152RR 275EL
A*34:02	19634	62RNR 62RR 62RRN 63NN 66NAQ 66NV 71QS 97I 113YR 145RT 245AS	62RNR 63NN 66NAQ 66NV 71QS 97I 113YR 145RT 245AS
A*33:01	19567	9T 62RNR 62RR 62RRN 63NN 66NAH 66NV 73ID 113YQ 186R 245AS	9T 62RNR 63NN 66NAH 66NV 73ID 113YQ 186R 245AS
A*66:02	19199	62RNR 62RR 62RRN 63NN 66NAQ 66NV 71QS 113YQ 145RT 245AS	62RNR 63NN 66NAQ 66NV 71QS 113YQ 145RT 245AS
A*33:03	19183	9T 62RNR 62RR 62RRN 63NN 66NAH 66NV 73ID 113YQ 186R 245AS	9T 62RNR 63NN 66NAH 66NV 73ID 113YQ 186R 245AS
A*11:01	19138	62QE 62QE+151H 62QE+s56G 66NAQ 66NV 71QS 97I 113YR 151AHA 152HA 163R 163RW 275EL	62QE 62QE+151H 62QE+s56G 66NAQ 66NV 71QS 97I 113YR 151AHA 152HA 163R 163RW 275EL
A*66:01	19102	62RNR 62RR 62RRN 63NN 66NAQ 66NV 71QS 113YQ 145RT 163R 163RW 245AS	62RNR 63NN 66NAQ 66NV 71QS 113YQ 145RT 163R 163RW 245AS
A*11:02	19088	62QE 62QE+151H 62QE+s56G 66NAQ 66NV 71QS 97I 113YR 151AHA 152HA 163R 163RW 275EL	62QE 62QE+151H 62QE+s56G 66NAQ 66NV 71QS 97I 113YR 151AHA 152HA 163R 163RW 275EL
A*68:01	19027	62RNR 62RR 62RRN 63NN 66NAQ 66NV 71QS 113YR 245VA	62RNR 63NN 66NAQ 66NV 71QS 113YR 245VA
A*25:01	19021	62RNR 62RR 62RRN 63NN 66NAH 66NV 76ESI 113YQ 145RT 163R 163RW 245AS	62RNR 63NN 66NAH 66NV 76ESI 113YQ 145RT 163R 163RW 245AS
A*29:02	18835	9T 62LQ 66NAQ 66NV 71QS 73TDA 76ANT 77NGT 113YR 245AS	9T 62LQ 66NAQ 66NV 71QS 73TDA 76ANT 77NGT 113YR 245AS
A*69:01	18063	62RNR 62RR 62RRN 63NN 66NAQ 66NV 71QS	62RNR 63NN 66NAQ 66NV 71QS
A*26:01	17195	62RNR 62RR 62RRN 63NN 66NAH 66NV 73TDA 76ANT 77NGT 113YQ 145RT 163R 163RW 245AS	62RNR 63NN 66NAH 66NV 73TDA 76ANT 77NGT 113YQ 145RT 163R 163RW 245AS
A*43:01	14771	62LQ 66NAH 66NV 73TDA 76ANT 77NGT 113YQ 145RT 163R 163RW 245AS	62LQ 66NAH 66NV 73TDA 76ANT 77NGT 113YQ 145RT 163R 163RW 245AS
A*80:01	12467	56E4 66NAH 66NV 76ANT 77NGT 97I 113YR 152RR	56E4 66NAH 66NV 76ANT 77NGT 97I 113YR 152RR
A*01:01	4440	44KM3 62QE 62QE+151H 62QE+s56G 66NAH 73TDA 76ANT 77NGT 97I 113YR 152HA 163R 163RG 275EL	44KM3 62QE 62QE+151H 62QE+s56G 66NAH 73TDA 76ANT 77NGT 97I 113YR 152HA 163R 163RG 275EL

A*36:01	4382	44KM3 62QE 62QE+151H 62QE+s56G 66NAH 73TDA 76ANT 77NGT 97I 113YR 152HA 275EL	44KM3 62QE 62QE+151H 62QE+s56G 66NAH 73TDA 76ANT 77NGT 97I 113YR 152HA 275EL
A*34:01	4015	62RNR 62RR 66RKQ 71QS 113YQ 145RT 245AS	62RNR 66RKQ 71QS 113YQ 145RT 245AS
B*15:16	2495	62RER 62RR 62RRN	
B*44:03	2071	None	
Patient 3 (cPRA 54.23%)			
B*37:01	12794	None	None
B*57:01	11036	62GE 62GRN 97V	62GE 62GRN 97V
B*57:03	10812	62GE 62GRN 97V	62GE 62GRN 97V
B*58:01	8634	62GE 62GRN	62GE 62GRN
A*02:03	7274	62GE 62GK2 145HT	62GE 62GK2 145HT
A*02:01	7104	62GE 62GK2	62GE 62GK2
A*02:06	6743	62GE 62GK2	62GE 62GK2
B*08:01	2210	None	

Highlighted epitopes compose the minimum epitope set that can explain the reactions observed with each SAB HLA antigen.

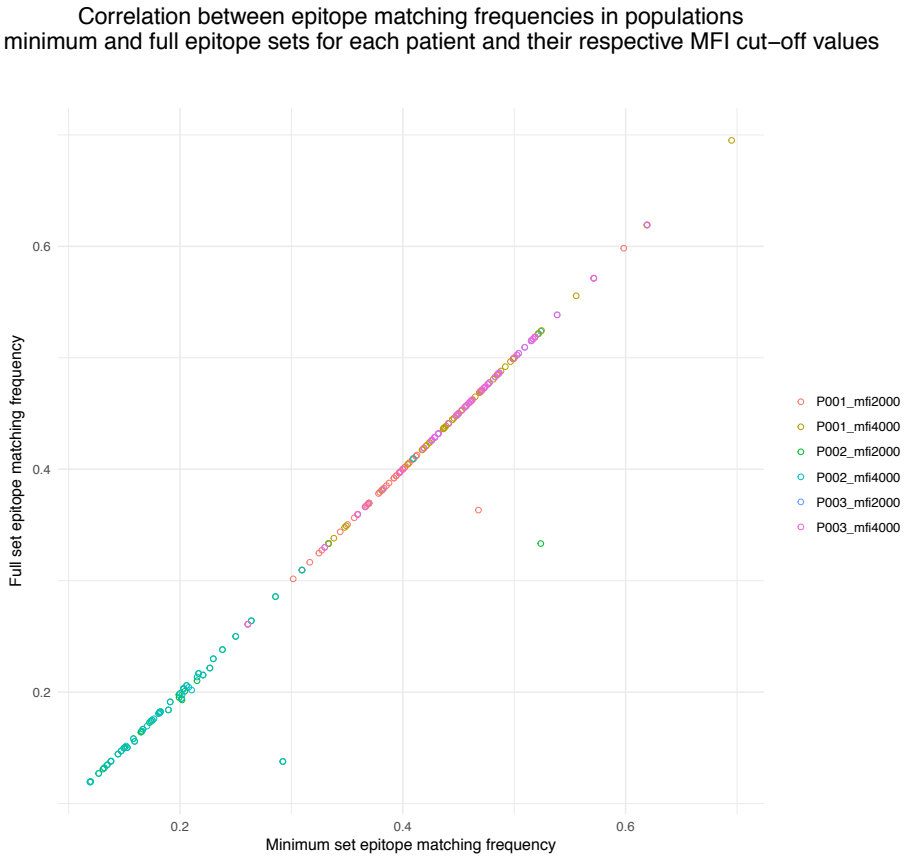


Figure 3.13: Correlation between epitope matching percentages in populations obtained using 'minimum' and 'full' epitope sets as input. The epitopes within those sets are not present in the respective patients (non-self). According to their SAB assay results and the HLA epitope logic, anti-HLA antibodies present in patients could be targeting some or all epitopes in the 'full' set. Thus, matched donors are the ones not having any of the epitopes in the analysed epitope set ('negative' matching). While the 'full' epitope set contains all non-self epitopes present in reactive HLA alleles, the 'minimum' epitope set contain the minimum set of non-self epitopes that could explain the observed reactivity. Each point in the plot represents the frequencies of individuals in a population not having the input epitope sets ('minimum' set: x-axis, 'full' set: y-axis) that could be donor candidates for each patient (three patients at MFI cut-offs of 2000 and 4000). Although it is expected by probability that smaller sets would yield higher 'negative' matching percentages, the plot shows that for the vast majority of the populations there was no difference in matching 'minimum' and 'full' epitope sets.

Comparison between population matching percentages obtained using allele and epitope sets generated by EpVix as input is shown in Figures 3.15 and 3.16. It can be noted that for most patients and the respective MFI cut-offs used, frequencies of matched individuals across populations are higher for alleles than epitopes, except for patient 3,

but only when combined to a 2000 cut-off value. Increasing the MFI cut-off removes the B*08:01 allele from the allele input set, for which the EpVix algorithm did not identify any possible reactive epitope. Higher allele matching frequencies in some populations for some patient / MFI cut-off combinations can be explained by the fact that HLA alleles included in SAB assays correspond to a subset of the existing alleles in populations, whereas alleles not included in those assays contain shared epitopes identified by EpVix.

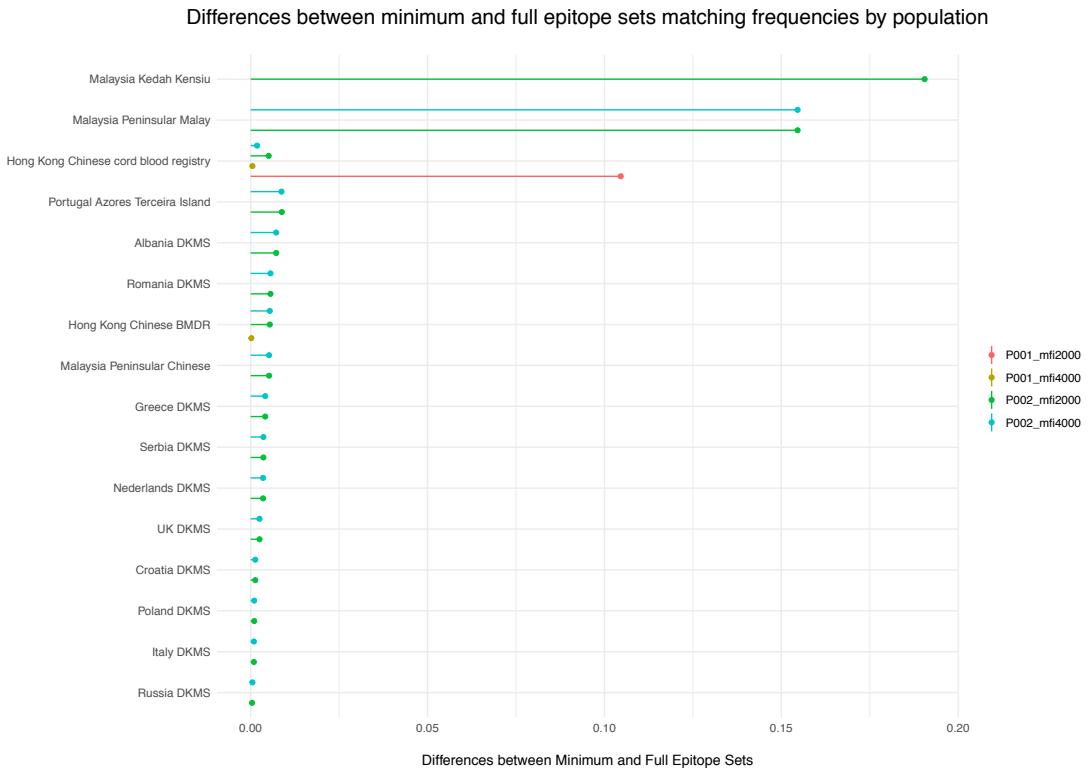


Figure 3.14: Dotchart of populations showing any difference in matching likelihoods between minimum and full epitope sets. Most differences are lower than 2%, except for one Chinese population and Malaysian populations showing differences higher than 10% for patient 1 and 2, respectively.

To ascertain the explanation for the aforementioned results, the same previously described matching analysis was performed including HLA alleles not part of Luminex® SAB assays that contain reactive epitopes identified (Figures 3.15 and 3.16). The higher allele matching frequencies previously identified were then eliminated upon the inclusion of non-SAB alleles, confirming our predictions. This result also highlights that alleles present in SAB assays may not sufficiently represent allele diversity in populations (Figure

3.17). Inclusion of non-SAB alleles also generates a slight increase in the quantity of epitope matching frequencies higher than allele matching frequencies. These differences are due to some alleles above the declared MFI cut-off not having any associated epitope defined as reactive by the HLA Matchmaker algorithm (Table 3.2).

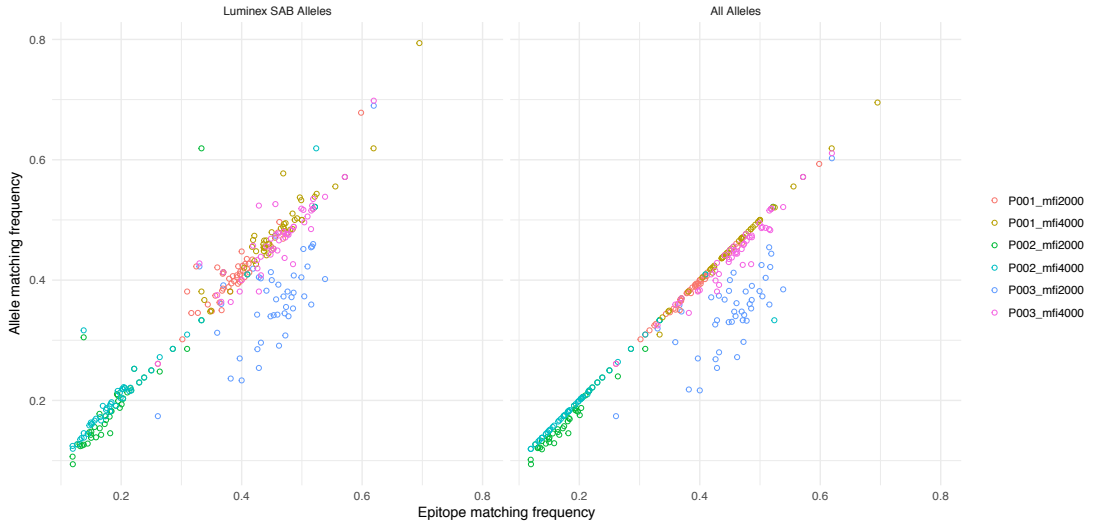


Figure 3.15: Correlations between population matching percentages obtained using allele and epitope sets as input. Each point in the plot represents the frequencies of individuals in the population not having the input sets (epitope set: x-axis, allele set: y-axis) in a population, for one of the three patients at a given MFI cut-off (2000 or 4000). In the panel on the left, alleles from input set include only Luminex® SAB alleles, while panel on the right includes alleles not present in the assay.

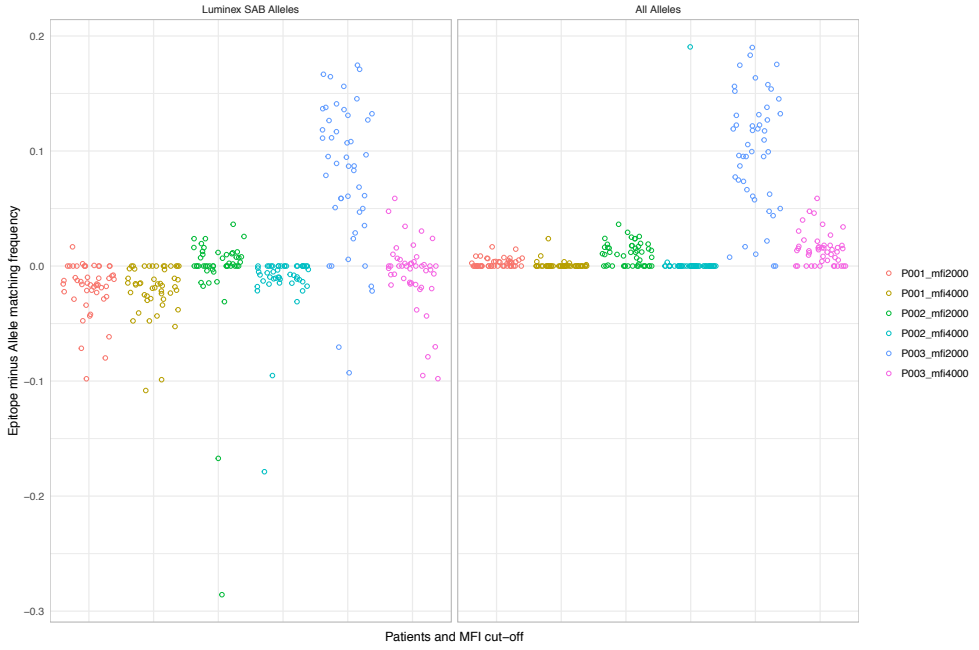


Figure 3.16: Point plot of epitope matching frequencies minus allele matching frequencies in populations, categorized by patients / MFI cut-off combinations. In the panel on the left, alleles from input set include only Luminex® SAB alleles, while panel on the right includes alleles not present in the assay.

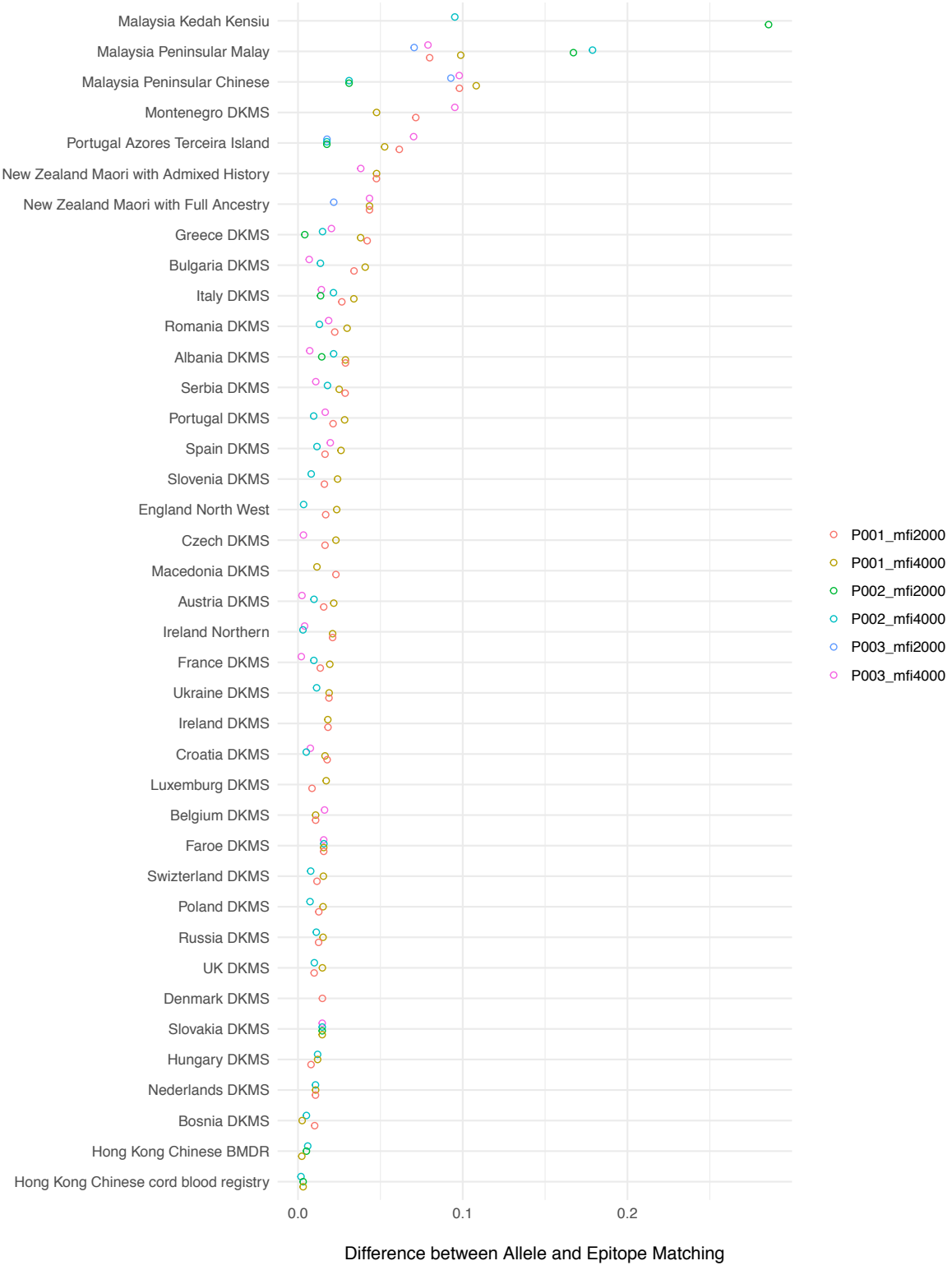


Figure 3.17: Dot chart showing only populations where Luminex® allele matching frequencies are higher than epitope matching frequencies. For these populations, SAB assays do not sufficiently represent their allele diversity, with higher discrepancies being observed in Malaysian populations.

The reason why some alleles in EpVix output show no related epitope identified as reactive is due to the logic of its algorithm, where self-epitopes or epitopes belonging to any allele below the MFI cut-off value are not flagged as reactive epitopes. This is the case for B*08:01 and other alleles found reactive in SAB assays, but more pronounced differences were observed due for B*08:01 in patient 3 at a 2000 cut-off due to its high frequency in the populations investigated [58]. Therefore, inconsistencies in HLA epitope definitions and / or SAB assays could be limiting factors in the optimization of transplantation matching towards increasing chances of finding a match. The difficult interpretation of MFI values along with several factors that can lead to false-positive or false-negative results are a source of controversy on the reliability of SAB assays [116,117,131].

A further challenge in the interpretation of HLA epitope data is that the logic behind HLA Matchmaker epitopes defined *in silico* may not completely reflect the *in vivo* epitope interactions with alloantibodies. HLA Matchmaker reports on the knowledge of HLA polymorphisms, and general definitions regarding what is known about antibody-antigen interactions, to infer HLA epitopes that *could* be reactive in mismatched transplanted organs. Polymorphic residues exposed on the HLA surface within an area consisting of a few residues that are believed to bind CDR-H3, and to be centrally located within the epitope (so-called functional epitope), are considered the relevant units ('eplets') to be compared between donors and recipients. On the other hand, traditional *in silico* B cell epitope prediction tools, mainly used in vaccine development, utilize different approaches to infer epitopes. These tools take into account physicochemical properties such as hydrophilicity, flexibility, polarity, and exposed surface, and spatial characteristics of the 3D structure of proteins to determinate surface accessibility [132]. Despite extensive research in this field, *in silico* B cell epitope prediction methods are far from reliable and are currently an unsolved problem [133-135], indicating that antibody-antigen interactions are much more complex than what it is currently known and that inferences used by HLA Matchmaker algorithm may be oversimplifying the mechanism of alloreactivity.

It has been suggested that predicting B cell epitopes should be accompanied by knowledge of binding antibodies [134], and the latest prediction tools have been addressing this issue by incorporating antibody data related to its sequence and / or structure [136]. In this sense, HLA Matchmaker algorithm also misses specific antibody information, since the determination of molecular surface areas over the HLA molecule corresponding to the paratope of a patient's antibody is not based on human anti-HLA alloantibodies, but uses similar models available on PDB to infer that information [10].

Recent studies have been challenging the paradigm of epitope-paratope interaction through CDRs and the pivotal importance of CDR-H3. With the growing number of new antibodies being characterized, it has been shown that epitope-paratope interactions are not exclusively influenced by CDRs, and also that not always CDR-H3 plays a dominant role in the specificity of the interaction [134]. This represents another limitation in the HLA Matchmaker algorithm, since HLA epitopes are majorly defined based on CDR-H3 complementarity. Besides, it has been demonstrated that residues outside the area predicted to interface CDR-H3 are critical for the reactivity of some HLA epitopes [137].

Finally, HLA epitope matching differs from traditional HLA allele matching by generating subunits containing polymorphic residues from HLA molecules which are configured according to general definitions regarding epitope interactions with CDR-H3. However, since current understanding regarding antigen-antibody interactions have not been reliable for epitope prediction, the parameters used to define 'eplets' would also inherit this knowledge gap. Further studies characterizing antigen-antibody interactions in humoral HLA alloreactivity are necessary for improving prediction of HLA epitopes.

3.5 Conclusion

The HLA Matchmaker algorithm has been setting the standards for HLA epitopes matching based on the structural concept of a functional epitope containing relevant polymorphisms that can potentially elicit an antibody immunological response. With the increasing interest in this field, EpFreq-DB provides a public database with HLA epitope frequency data for several worldwide populations. Nevertheless, since a new release of HLA epitope registry has changed some HLA epitope definitions, EpFreq-DB is being updated accordingly and an upgrade will be released at a later stage. Worldwide epitope distribution described in the present study shows that several epitopes are almost always

present or absent across all investigated populations, narrowing a subset of epitopes with higher variability to be more relevant for transplant matching.

A comparative analysis of alloreactivity profile matching using HLA epitopes and high resolution HLA allele matching highlights current limitations in both SAB assays and HLA epitope definitions. It should be noted that this analysis is limited by a small patient sample size, and additional investigations using a larger number of patients are needed to confirm the present findings.

Chapter 4

Structural basis of the association of HLA-C*04:01 with nevirapine adverse drug reactions

4.1 Abstract

HIV patients being treated with nevirapine are at risk of developing adverse drug reactions, ranging from mild to severe hepatic and skin reactions. These T cell mediated 'off-target' reactions have been shown to be associated with certain human leukocyte antigen (HLA) alleles in patients being treated with particular drugs. Previous studies have reported associations of HLA-C*04:01 with nevirapine hypersensitivity in African populations, including a genome-wide association study (GWAS) reporting an association to a SNP encoding a substitution at residue Glu49 of HLA-C*04:01. A putative role of this residue in causing nevirapine-induced hypersensitivity had been suggested due to the fact that this residue is nearly unique to HLA-C*04 alleles. To investigate this hypothesis, molecular docking of nevirapine and 12-hydroxy-nevirapine metabolite (a break-down product of nevirapine that could be involved in the adverse reaction) with HLA-C*04:01 was undertaken to determine the possible effect of the residue substitution on nevirapine binding. The data suggest that none of the predicted modes of nevirapine docking conformation interact with residue Glu49, which is located on the periphery of the peptide binding domain. The putative docking site with the highest predicted affinity highlights an interaction between nevirapine and residues Ser9 and Phe99 in the B pocket, which are almost unique to C*04:01, with the exception of relatively rare alleles C*04:07 and C*14:02. While none of the predicted modes interacted with residue 49, docking results suggest that binding of either nevirapine or 12-hydroxy-nevirapine around the centre of the peptide-binding regions is likely to be important in the mechanism of the immune-mediated reaction.

4.2 Introduction

Nevirapine (Figure 4.1) is an anti-retroviral drug used in the treatment of human immunodeficiency virus 1 (HIV-1) infected patients. Commercialized as Viramune®, it is part of the non-nucleoside reverse transcriptase inhibitor (NNRTI) type of anti-retroviral drugs [138]. Listed in the WHO Model List of Essential Medicines containing essential drugs needed for any basic health-care system [139], nevirapine has been shown to have long-term effectiveness [140] and prevent vertical transmission [141]. It is administered in highly active antiretroviral therapy (HAART) regimens, consisting of a combination of three or more drugs. These regimens are intended to control the disease by providing a sustained suppression of HIV-1 replication, lowering viral load and increasing CD4+ T-cell levels [142].

After invading a host cell, HIV and other single-strand RNA retroviruses use the reverse transcriptase (RT) enzyme to reverse-transcribe its RNA genome into double-stranded DNA, which is then integrated into the host cell DNA in the nucleus, allowing the virus to use the cell machinery for its replication. Nevirapine inhibits RT by an allosteric binding to the hydrophobic pocket of its p66 subunit, close to the catalytic site [143]. Figure 4.2 shows interaction of HIV-1 RT with a representative NNRTI (UC781), to exemplify nevirapine mechanism of action [144].

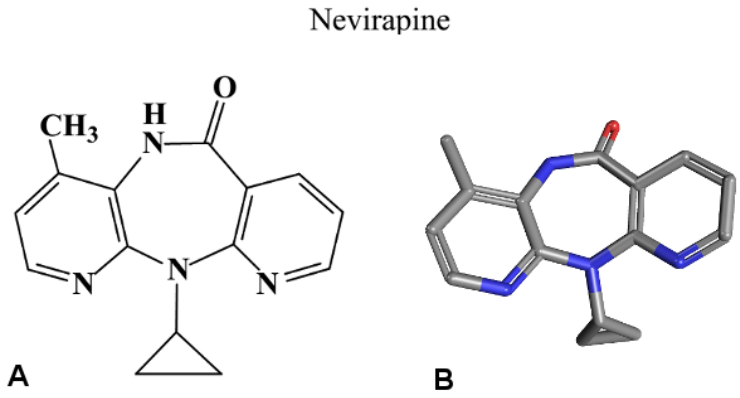


Figure 4.1: Nevirapine 2D (A) and 3D (B) chemical.

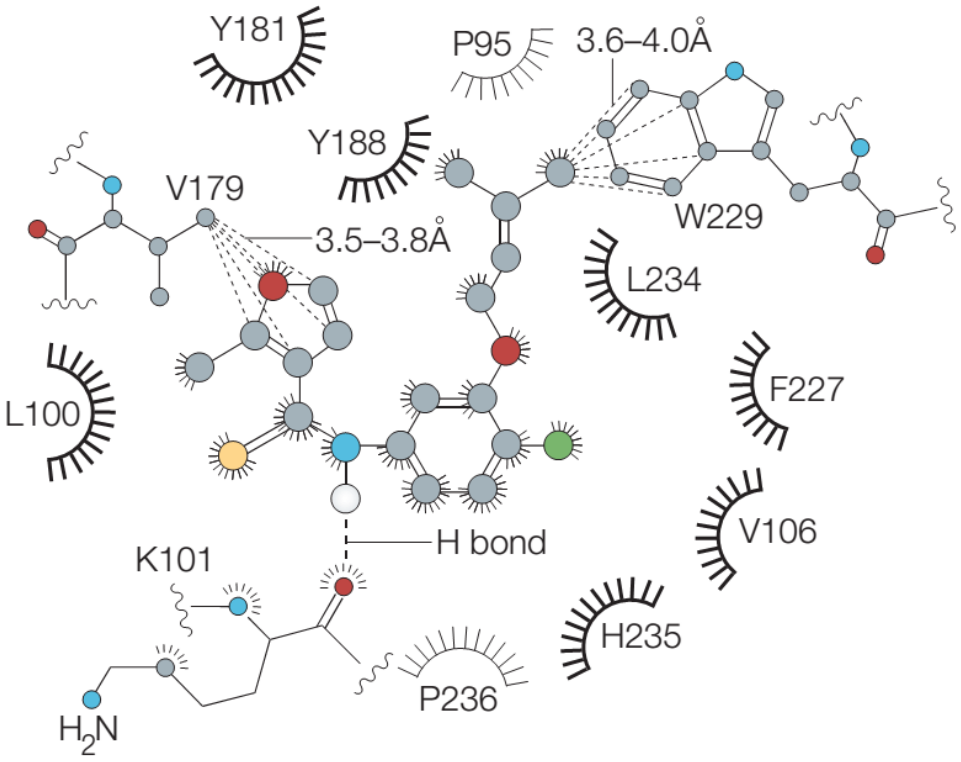


Figure 4.2: A model for the interaction of HIV-1 RT with a representative NNRTI (UC781), to exemplify the functional interaction of HIV-1 RT and nevirapine. The hydrogen bond with K101, and the two methyl-group-aromatic-ring interactions are shown explicitly. Other main hydrophobic contacts are shown with bold lines; minor ones are shown with faint lines [144].

In Sub-Saharan Africa, a region with the highest worldwide incidence of HIV [145], nevirapine-based regimens have been a therapy of choice for fighting against the virus due to its lower cost in comparison to other anti-retroviral drugs with comparable efficacy [146], and to its ability to prevent vertical transmission, reducing the prevalence of HIV in the next generations [147]. A meta-analysis of randomized trials for HIV-1 anti-retroviral drugs had shown that the efficacy of nevirapine regimens is similar compared to non-nevirapine regimens [148]. However, nevirapine has showed a higher risk for discontinuation of treatment due to adverse drug reactions when compared to other anti-retroviral drug types, such as ritonavir-boosted protease inhibitors [149]. Long-term adherence to the treatment is crucial for the success of the treatment regimen, thus it is important that adverse reactions are minimised to avoid treatment discontinuation [138].

Generally, ADRs are classified as 'on target' type A and 'off target' or 'idiosyncratic' type B. The latter have unpredictable nature with unclear relation to the drug's

pharmacological mechanism, differing than more predictable dose-dependent type A reactions displaying symptoms related to the drug's known pharmacological mechanism [150]. Most nevirapine-induced ADRs are skin hypersensitivity and hepatic reactions belonging to a type B subcategory. This subtype of reaction is characterized by delayed-type hypersensitivity reactions mediated by T cell response, which can be systemic or organ specific [151]. Mild and severe hepatic reactions known as drug-induced liver injury (DILI) accounts for around 4% of nevirapine-induced ADRs [152], manifesting as hepatocellular injury or cholestasis (obstruction of normal bile flow) that can progress to fulminant hepatic failure [153].

Skin hypersensitivity configures as systemic reactions accounting for most nevirapine ADRs, where clinical trials reported 17% of the patients developing skin rashes, with 0.3% showing severe skin reactions, such as Steven Johnson Syndrome (SJS) and Toxic Epidermal Necrolysis (TEN) [150]. SJS and TEN are the most severe bullous skin ADR, characterized by a macular purple-red exanthema that can be painful, and it can develop quickly after its establishment. They usually appear during the first 6 weeks of treatment and stopping the treatment may not improve the symptoms, but only prevent them from continuing to develop. Currently SJS and TEN are considered to be the milder form and the more severe form of the same disease, respectively, where SJS is defined by less than 10% of skin detachment while TEN defines more than 30% of skin detachment. Intermediate forms between 10-30% of skin detachment are considered SJS/TEN overlap [154]. SJS has 13% mortality rate and TEN has a 39% mortality rate, while SJS/TEN has a 21% mortality rate [155].

Although the functional mechanisms of those ADRs are not yet fully understood, genetic variability in patients has been observed to influence their incidence in different geographical regions with distinct ethnic makeup [156]. Probably due to T cell involvement in those reactions, several HLA polymorphisms have been found associated with type IV ADRs from various drug classes [52].

Nevirapine-induced hypersensitivity reactions have been associated with many class I and class II alleles in different populations, but HLA-C*04:01 seems to be a risk factor common to populations with different ethnicities [157]. Nevirapine-induced SJS/TEN has been found associated with HLA-C*04:01 in genome-wide association studies (GWAS) using a Malawian discovery cohort and a replication cohort from three countries (Malawi, Uganda and Mozambique) [158,159]. Other HLA and ADR associations include

the well known associations of HLA-B*57:01 and hypersensitivity to abacavir anti-RT drug, HLA-B*15:02 and carbamazepine-induced SJS/TEN and HLA-B*58:01 and allopurinol hypersensitivities [160].

To understand the mechanisms of those associations on a structural level, computational methods such as molecular docking have been applied to predict the binding mode and affinity between drugs (or metabolites) and HLA molecules. This method contributed to a thorough characterization of the interaction between abacavir and HLA-B*57:01, revealing a strong binding of abacavir to the peptide binding groove of the HLA molecule resulting in presentation of a modified peptide repertoire [161].

Regarding the underlying mechanism of HLA associations with adverse drug reactions, three main hypotheses have been suggested: the hapten, the p-i concept and the altered peptide repertoire. The hapten model states that drugs can covalently bind to proteins (hapten-carrier complexes) which undergo antigen processing and presentation, being able induce a novel immune response. The p-i concept (or pharmacological interaction of a drug with immune receptors) postulates that a drug might directly bind to TCR and stimulate the T cell. The altered peptide repertoire model (demonstrated for abacavir interaction with B*57:01) results from a non-covalent binding of the drug within the HLA antigen-binding cleft, modifying the shape of the antigen-binding cleft leading to a consequent altering of the presented peptide repertoire [155,162]. The peptides presented are thus not recognised as "self", giving rise to an auto-immune like reaction. Given that the altered peptide repertoire model is the only one currently supported by structural data, the research presented in this chapter is working under the assumption that the altered peptide repertoire is also likely to be the mechanism of action underlying adverse reactions to nevirapine.

In the aforementioned GWAS of nevirapine treated Malawian patients, the SNP (rs1050409) found associated encodes for a substitution of alanine to glutamic acid at residue 73 of the full length HLA-C*04:01 sequence (corresponds to position 49 when excluding the signal peptide). Additionally, this substitution is nearly unique to HLA-C*04 alleles, suggesting a putative role in nevirapine-induced ADR mechanisms. Since HLA protein sequences found in relevant immunogenetic and protein databases consider the polypeptide of the mature protein for sequence positions, i.e. excluding the signal peptide, position 49 instead of 73 is going to be used throughout this chapter to indicate the substitution.

In collaboration with Prof Munir Pirmohamed's research group at the University of Liverpool, the present chapter describes the investigation of the structural basis of association of HLA-C*04:01 with nevirapine-induced ADRs using molecular docking methods. The putative role of Glu49 residue in HLA-C*04:01 (and other Glu49 HLA-C*04 alleles) in inducing ADRs in patients under nevirapine treatment was a starting point for the present investigations. Docking analyses were also performed using a the 12-hydroxy-nevirapine (a breakdown metabolite derived from nevirapine) since studies using in rats suggest that nevirapine-induced skin rash is associated to its metabolite 12-hydroxy-nevirapine [163].

The research described in this chapter has been published as part of a larger study regarding genetic associations with nevirapine hypersensitivity in the Journal of Antimicrobial Chemotherapy (2017) [159], corresponding to my contribution in the manuscript (Appendix A).

4.3 Methods

4.3.1 HLA-C*04:01 crystal structure

A search for the crystal structure of HLA-C*04:01 or other closer alleles was performed by querying UNIPROT database (http://www.uniprot.org/) [164] for "hla-c" AND organism: "Homo sapiens (Human) [9606]", resulting in 42 reviewed hits. The 'P30504' UNIPROT ID, described as HLA Cw-4 alpha chain (MHC class I antigen Cw*4), was queried in Protein Data Bank (PDB) (http://www.rcsb.org/pdb) [165], retrieving the 1QQD PDB crystal structure [166], described as HLA-Cw4, which is a nomenclature referring to the antigenic classification. By mapping its protein sequence with HLA protein sequences in IMGT/HLA database [107], 1QQD was identified as HLA-C*04:01 crystal structure.

4.3.2 Choice of control HLA molecules

Assuming residue Glu49 as a putative binding site for nevirapine in C*04:01, a control molecule for comparison of binding affinity should have an alternative residue at 49

position. To choose the most similar molecule to C*04:01, but containing other residue than a glutamic acid in position 49, the following steps were performed: i) HLA-C protein sequences were extracted from IMGT/HLA database, selecting only protein sequences containing a residue other than glutamic acid in position 49, ii) *makeblastdb* application, which is part of the *blast+* package from NCBI (ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/) [167] was used to build a blast database containing only the selected sequences, iii) the C*04:01 FASTA sequence from 1QQD PDB entry was queried against the built database using *blastp* application, also part of *blast+* package. C*04:07 was the top hit in *blastp* results, containing only one residue change as an alanine in position 49 (Ala49), thus selected as a control molecule. The next hit belonging to the C*04 allele group was C*04:03, which was also included as control molecule, containing six residue changes compared to C*04:01 (Tables 4.1 and 4.2). The next highest frequency HLA-C allele in the Malawi cohort (C*02:10) was also selected as a control molecule, under the assumption the genetic data indicated that allele was not associated with a nevirapine-ADR. To assess if Glu49 is a key residue for binding nevirapine independent of other residue differences in C*04:01 with other alleles, C*04:04 and C*04:16 were selected for nevirapine docking, as they have the same residue as C*04:01 in position 49 (Glu49) and are not rare alleles in world populations [58]. For another hypothesis that the binding site is located elsewhere, other HLA-C alleles present in Malawi cohort were selected as control molecules (C*02:02, C*03:02, C*03:03, C*03:04, C*06:02, C*07:01, C*07:02, C*07:04, C*08:02, C*12:03, C*14:02, C*15:05, C*16:01, C*17:01, C*18:01).

4.3.3 Modelling of control HLA molecules

Since protein crystal structures for control HLA molecules were not available at PDB, homology modelling was performed to generate their structures using two applications from Rosetta modelling software (version 2015.05). First, comparative modelling of HLA alleles using C*04:01 as a template was performed using the RosettaCM application [168]. This application uses two input files: the FASTA sequence of the molecule to be modelled and an alignment with the template sequence. Alignments were in HHR format, which were generated using *hhalign* application from HHsuite version 2.0 [169]. Another Rosetta application named *relax* [170] was also used to generate structure models of HLA alleles. This application uses an edited PDB file as an input file. In the present study, edited

versions of 1QQD PDB file were used for each HLA molecules to be modelled, where residue changes were manually altered by changing the residue names and removing atoms other than N, C, CA and O. The *relax* application reads the edited PDB files and autocomplete the specified residues and changes the structure configuration accordingly. For each molecule and each method, ten models were generated.

4.3.4 Nevirapine Docking

The preparation of molecules for docking, i.e. generation of PDBQT files, was performed using AutoDock Tools (ADT) applications from MGL Tools version 1.5.6 [171], both Graphical User Interface (GUI) and python scripts for batch generation of PDBQT files (prepare_receptor4.py for rigid molecule and prepare_flexreceptor4.py for flexible molecules). This pre-docking step is performed to ensure that atoms are in the correct format for docking, by adding partial charges, merging non-polar hydrogens and detecting aromatic carbons. The PDBQT format is 'PDB' from the original format plus 'Q' for partial charges and 'T' for AutoDock atom type. For the ligand, it also defines its flexible bonds (torsions) which allows for testing different conformations of the ligand to fit in the receptor. ADT has identified only one torsion in nevirapine, since most of the molecule is composed by aromatic rings, which are not flexible (Figure 4.1). Docking was performed using both rigid receptor and flexible receptor. A flexible receptor PDBQT file contains information of selected residues defined as flexible, i.e. their side chains can present different conformations, providing different shapes of binding sites to which the ligand can bind to. To determine which residues are flexible, all the models generated for each HLA molecule by Rosetta were superimposed in PyMOL Molecular Graphics System (www.pymol.org) [172]. Residues showing alternate side chains positions when comparing all the models were defined as flexible. This information was then used in ADT to generate a flexible residues PDBQT file. Autodock VINA [173] was used to perform nevirapine docking in HLA molecules for both rigid and flexible docking using the following parameters: a search space, i.e. the receptor area to search for likely binding sites, of 42 x 50 x 50 Å and exhaustiveness = 30. Visualisation and figures were produced in PyMOL and graphs were produced in the R programming language [89].

Table 4.1: HLA polymorphisms selected for nevirapine docking, including the C*04:01 and other HLA control molecules.

Allele	Description
C*04:01	Associated with nevirapine-induced adverse drug reaction in the Malawi cohort, including a SNP within the peptide sequence encoding a Glu49 substitution [158,159]
C*04:07, C*04:03	C*04 top hits on blastp of C*04:01 against other HLA-C alleles non-Glu49. C*04:07 is also present in the Malawi cohort.
C*04:04, C*04:16	Frequent alleles containing Glu49 but differing in other residues from C*04:01
C*02:02, C*02:10, C*03:02, C*03:03, C*03:04, C*06:02, C*07:01, C*07:02, C*07:04, C*08:02, C*12:03, C*14:02, C*15:05, C*16:01, C*17:01, C*18:01	Other HLA-C alleles present in Malawi cohort

Table 4.2: Polymorphic residue difference between C*04:01 and chosen control alleles within the peptide binding region protein sequence.

Allele	1	9	11	14	16	21	24	35	49	66	73	77	80	90	91	94	95	97	99	103	113	114	116	138	143	147	152	156	163	170	173	177
C*04:01	**G**	**S**	**S**	**W**	**G**	**R**	**A**	**R**	**E**	**K**	**A**	**N**	**K**	**D**	**G**	**T**	**L**	**R**	**F**	**L**	**Y**	**N**	**F**	**T**	**T**	**W**	**E**	**R**	**T**	**R**	**E**	**E**
C*04:07	*	-	-	-	-	-	-	-	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
C*04:03	*	Y	A	R	S	H	-	-	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
C*04:04	*	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	L	-	-	-	-
C*04:16	-	Y	A	R	-	H	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
C*02:10	C	Y	A	R	S	H	-	-	A	-	T	-	-	A	-	-	-	-	Y	-	-	D	S	-	-	-	-	W	E	-	-	-
C*02:02	C	Y	A	R	S	H	-	-	A	-	T	-	-	A	-	-	-	-	Y	-	-	D	S	-	-	-	-	W	E	-	-	-
C*03:02	-	Y	A	R	-	H	-	-	A	-	T	S	N	A	-	I	-	-	Y	V	-	D	S	-	-	-	-	L	L	-	K	-
C*03:03	-	Y	A	R	-	H	-	-	A	-	T	S	N	A	R	I	I	-	Y	V	-	D	Y	-	-	-	-	L	L	-	K	-
C*03:04	-	Y	A	R	-	H	-	-	A	-	T	S	N	A	-	I	I	-	Y	V	-	D	Y	-	-	-	-	L	L	-	K	-
C*06:02	C	D	A	R	-	-	S	-	A	-	-	-	-	-	-	-	-	W	Y	-	-	D	S	-	-	-	-	W	-	-	-	-
C*07:01	C	D	A	R	-	-	S	-	A	N	-	S	N	-	-	-	-	-	Y	-	-	D	S	-	-	L	A	L	-	-	-	-
C*07:02	C	D	A	R	-	-	S	-	A	-	-	S	N	-	-	-	-	-	S	-	-	D	S	-	-	L	A	L	-	-	-	-
C*07:04	C	D	A	R	-	-	S	-	A	-	-	S	N	-	-	-	F	-	Y	-	-	D	-	-	-	L	A	D	-	-	-	K
C*08:02	C	Y	A	R	-	-	-	Q	A	-	T	S	N	A	-	-	-	-	Y	-	-	-	-	K	-	-	-	-	-	-	-	K
C*12:03	C	Y	A	R	-	-	-	-	A	-	-	S	N	A	-	-	-	W	Y	-	-	D	S	-	-	-	-	W	-	-	-	-
C*14:02	C	-	-	R	-	-	-	-	A	-	T	S	N	A	-	-	-	W	-	-	-	D	S	-	-	-	-	-	-	-	-	-
C*15:05	C	Y	A	R	-	H	-	-	A	N	T	-	-	A	-	I	I	-	Y	-	H	D	-	-	-	-	-	L	-	-	-	-
C*16:01	C	Y	A	R	-	-	-	-	A	-	T	S	N	A	-	-	-	W	Y	-	-	D	S	-	-	-	A	Q	-	-	-	-
C*17:01	-	Y	A	R	-	-	-	-	A	-	-	-	-	A	-	-	I	-	Y	-	-	-	-	-	S	L	-	L	E	G	-	-
C*18:01	C	D	A	R	-	-	S	-	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

'-' = Same residue as reference sequence C*04:01. '*' = Residue has not been sequenced.

4.4 Results and Discussion

A collaborating research group under the leadership of Prof Munir Pirmohammed performed a GWAS in a Malawi cohort of patients being treated with nevirapine, finding an association of nevirapine-induced ADR with a SNP encoding for glutamic acid in position 49 of the HLA-C*04:01 allele [159]. This residue is present in almost all C*04 alleles and seldom observed in other HLA-C alleles, and it has been suggested to have a structural involvement in the mechanism of nevirapine-induced ADR associated with C*04:01. It has also been suggested that not only C*04:01, but all other C*04 alleles containing glutamic acid in position 49 (corresponding to the majority of C*04 alleles) would be also associated to ADRs. Table 4.3 summarises results found by the mentioned research group, showing a highly significant association of C*04:01 and nevirapine-induced SJS/TEN.

Table 4.3: Comparison of HLA-C allele frequencies between controls and SJS/TEN patients in the Malawi cohort.

HLA Allele	Controls (%)	SJS/TEN (%)	OR	CI (95%)	P-Value
C*04:01	0.25 (39)	0.64 (23)	5.12	2.24 - 12.16	2.25×10^{-5}
C*02:10	0.24 (37)	0.11 (4)	0.39	0.09 - 1.22	ns
C*03:02	0.05 (8)	0.03 (1)	0.52	0.01 - 4.09	ns
C*03:03	0.04 (6)	0.03 (1)	0.70	0.01 - 6.07	ns
C*03:04	0.08 (12)	0.14 (5)	1.89	0.49 - 6.29	ns
C*06:02	0.24 (36)	0.08 (3)	0.3	0.06 - 1.04	0.04
C*07:01	0.16 (24)	0.25 (9)	1.79	0.66 - 4.55	ns
C*07:02	0.08 (13)	0.06 (2)	0.63	0.07 - 3.01	ns
C*07:04	0.05 (7)	0.06 (2)	1.23	0.12 - 6.83	ns
C*08:02	0.13 (20)	0.11 (4)	0.83	0.19 - 2.73	ns
C*12:03	0.04 (6)	0.06 (2)	1.44	0.14 - 8.50	ns
C*16:01	0.1 (15)	0.08 (3)	0.84	0.15 - 3.21	ns
C*17:01	0.21 (32)	0.19 (7)	0.91	0.31 - 2.39	ns
C*18:01	0.14 (22)	0.08 (3)	0.54	0.1 - 1.98	ns

'HLA genotyping' refers to individuals from Malawi cohort whose HLA type has been confirmed by histogenetics, while 'Imputed HLA Alleles' comprises all individuals in the Malawi cohort whose HLA type has been estimated by imputation. HLA genotyping sample size: controls = 35, cases = 81; Imputed HLA Alleles sample size: controls = 182; cases = 151.

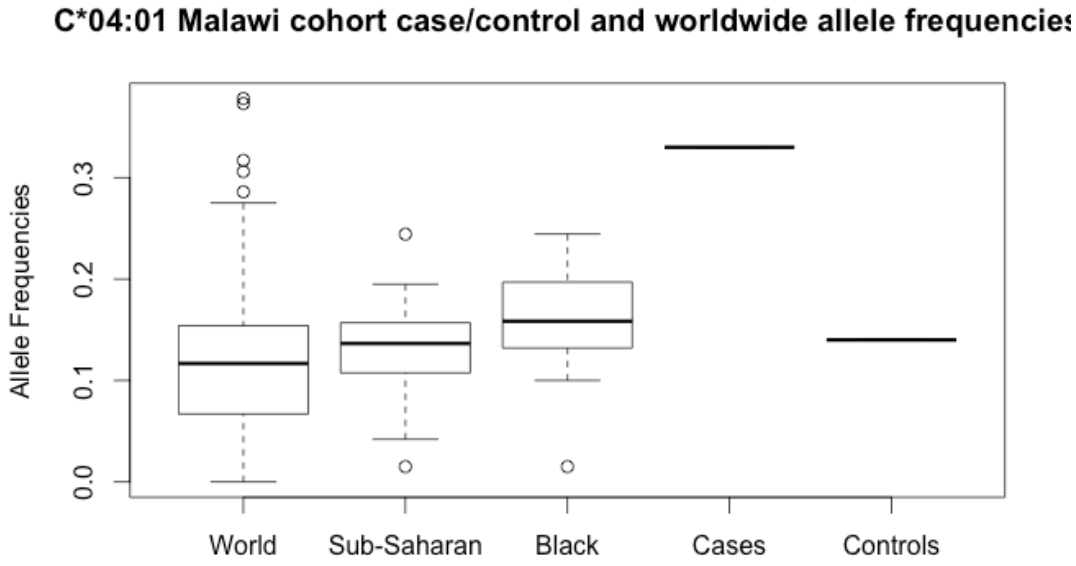


Figure 4.3: Comparison of C*04:01 frequencies observed in cases and controls from the Malawian cohort with frequencies from worldwide healthy populations from AFND ('World', 'Sub-Saharan', 'Black') [58].

Figure 4.3 shows that HLA-C*04:01 frequency from controls is similar to most frequencies observed in world populations, especially Sub-Saharan and populations defined as having 'Black' ethnicity. In contrast, frequency of this allele observed in patients fall out of the range observed for all healthy population, except from some outliers. This comparison reinforces the HLA-C*04:01 association with nevirapine-induced ADRs, highlighting that the chance of its frequency value to occur by chance in any world population is very low.

Computational docking of nevirapine in the HLA-C*04:01 peptide binding region was performed to predict the binding site of nevirapine where the total molecular area of the α1 and α2 regions was used as a search space. Figure 4.4 shows a total of 20 conformations predicted to bind C*04:01. Most of the conformations are within the peptide binding region, while others are restricted to a specific loop outside the peptide binding region. Glutamic acid in position 49 is highlighted in Figure 4.4, showing that no predicted conformations of nevirapine docking were able to contact this residue. The identification of the residues within a 4 Å radius of all the conformations predicted does not include position 49 within this area.

The same analysis was performed using 12-hydroxy-nevirapine as a ligand, which is a breakdown metabolite derived from nevirapine and suggested to be associated with ADR in comparison to other metabolites. Figure 4.5 shows 20 predicted conformations of 12-hydroxy-nevirapine binding C*04:01. In this case all conformations are within the peptide binding groove. Similar to nevirapine, no predicted conformation was able to contact glutamic acid in position 49. The region containing residue 49 does not seem to be reachable by nevirapine, since it is located in the bottom of the α1 and α2 molecular structure, far from the regions where nevirapine has been predicted to bind, which are the peptide binding groove (for both nevirapine and 12-hydroxy-nevirapine) and a cavity located at the top of the molecule outside near residue 14 (nevirapine only).

To identify the putative binding site of nevirapine or 12-hydroxy-nevirapine, residues around 4 Å of all the predicted modes were analyzed using PyMOL (Figure 4.6). From all residues identified within this radius area, positions where residues are polymorphic among the alleles in the Malawi cohort, where at least 50% of the alleles have a residue different than C*04:01 were selected as potential candidates (green columns in Table 4.4). By analyzing residue differences in those positions between alleles, candidates for ligand binding site can be identified. The three top positions where most of the other alleles in the cohort have different residues from C*04:01 are position 14, 9 and 99. Position 14 in C*04:01 shares the same residue only with C*04:07. Position 9 shares the same residue with C*04:07 and C*14:02. Position 99 shares the same residue with the same alleles as position 9 plus C*18:01. Other positions share residues with four or more alleles. Predominance of C*04:01 residues in those specific positions in comparison to other alleles in the cohort suggests their putative role in interaction with nevirapine or its metabolites.

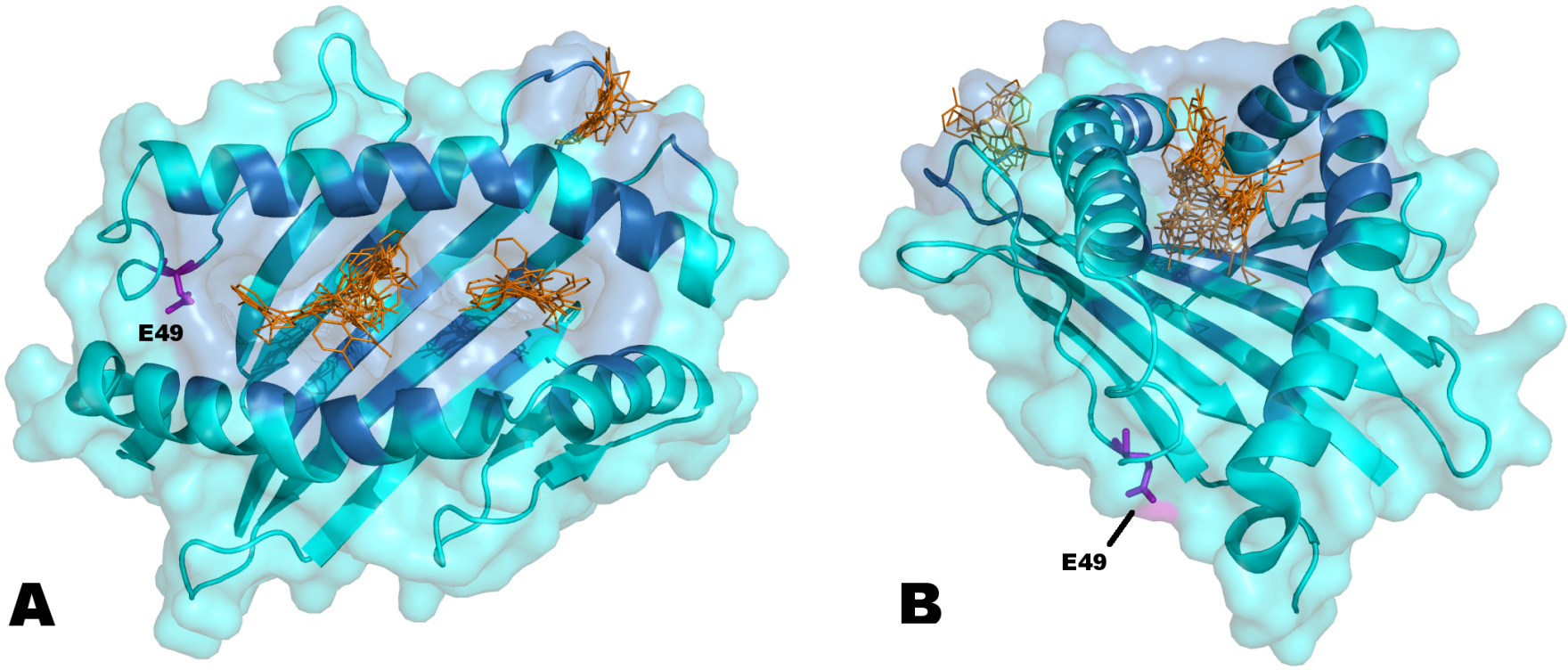


Figure 4.4: Docking of nevirapine to HLA-C*04:01 peptide binding region, where a total of 20 conformation modes were produced. 'A' shows the top view of the molecule, while 'B' shows a side view of the molecule. Although SNPs coding for residue position 49 in the molecule are strongly associated with ADR, no predicted conformation binds that position in the molecule. Areas coloured in darker blue are contain polymorphic residues contacting nevirapine conformations.

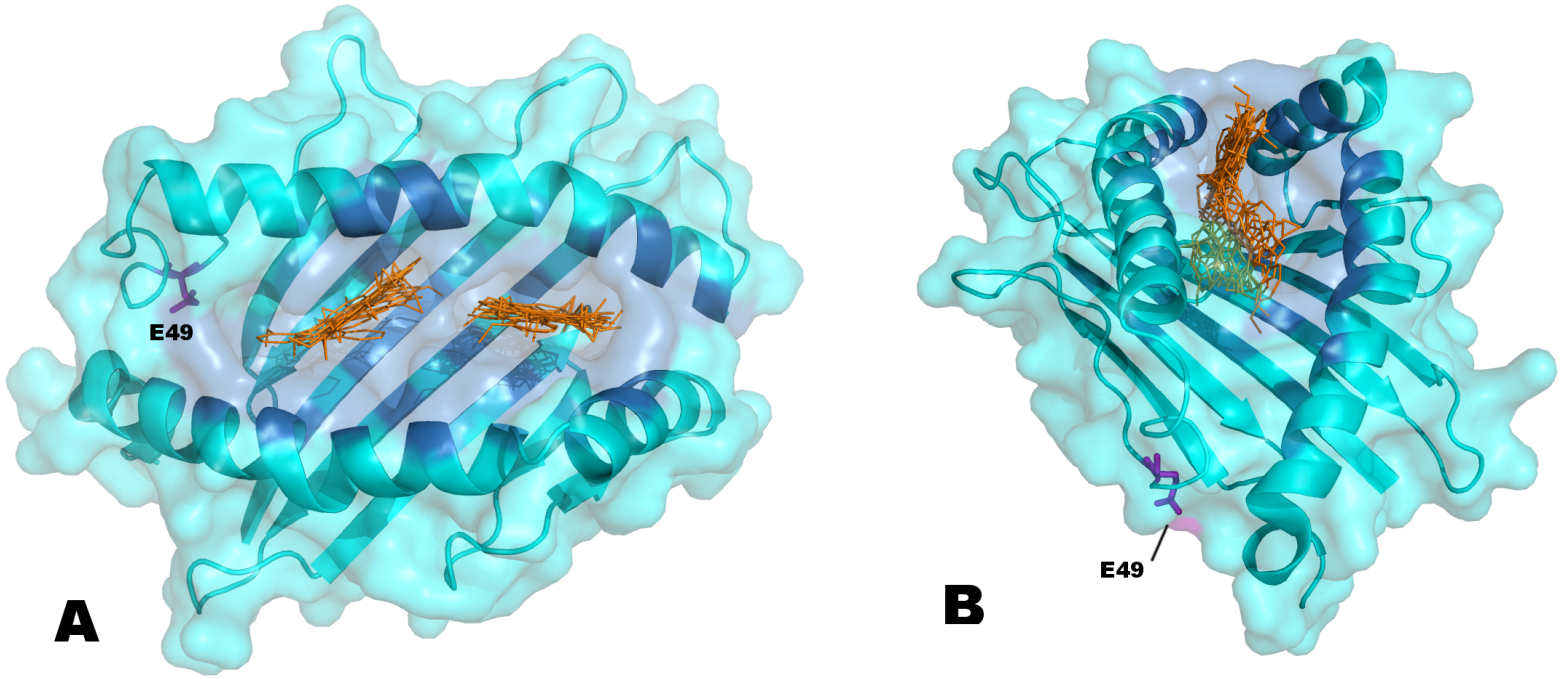


Figure 4.5: Docking of 12-hydroxy-nevirapine to HLA-C*04:01 peptide binding region, where a total of 20 conformation modes were produced. 'A' shows the top view of the molecule, while 'B' shows a side view of the molecule. 12-hydroxy-nevirapine is a nevirapine metabolite, which has been shown to be associated to ADR in comparison to other nevirapine metabolites. Although SNPs coding for residue position 49 in the molecule are strongly associated with ADR, no predicted conformation binds that position in the molecule. Areas coloured in darker blue are contain polymorphic residues contacting 12-hydroxy-nevirapine conformations.

Docking results also show that predicted conformations form clusters (3 clusters for nevirapine and 2 clusters for 12-hydroxy-nevirapine) which may be indicating the possible binding sites. While for nevirapine the conformations within each cluster are more spread and present variable orientations, for 12-hydroxy-nevirapine, the conformations seem to be limited to a smaller molecular area and follow a similar orientation. Figure 4.6 shows that for both nevirapine and 12-hydroxy-nevirapine one cluster is limited to a region that includes residues 9 and 99, and the other cluster interacts with several other polymorphisms where residue 156 contains more amino acid differences in other alleles not C*04:01. Another cluster specific to nevirapine analysis contacts position 14. Although this region is outside the peptide binding groove, position 14 is accessible for interaction with other proteins, since it is situated in the top of the α1 and α2 molecular structure.

Position 14 is also unique to C*04:01, with exception of C*04:07. In the Malawi cohort only one individual (patient) was typed for C*04:07 and the highest frequency that has been observed in Sub-Saharan African populations is 4% (Figure 4.9). Furthermore, most studies use low resolution HLA typing, where both C*04:01 and C*04:07 are part of Cw4 classification. Additionally, position 49 is the only residue difference between C*04:01 and C*04:07 mature protein sequences, which is unlikely to be a binding site for nevirapine and its metabolites, or to interact with other molecules of the immune system.

Each predicted conformation is assigned a score representing the energy of the docking between ligand and molecule for that conformation. The lowest scores are assumed to be the best conformations to fit the molecule, having the lowest free energy. However docking scores have been shown to be poor predictors of the best conformation [174]. Figure 4.7 shows the first poses (conformations with the lowest scores) from docking of nevirapine and 12-hydroxy-nevirapine and the contacting residues (conserved and polymorphic among alleles in Malawi cohort). Both are within the peptide binding region, but are in contact with different set of residues. Taking into account the polymorphic positions which are predominantly different than C*04:01 in comparison with other alleles in the Malawi cohort (Table 4.4), nevirapine binding site includes two positions nearly unique to C*04:01 (residues 9 and 99), while for 12-hydroxy-nevirapine first docked mode the position which has more differences from C*04:01 among Malawi cohort alleles is 156 (four alleles have the same residue as C*04:01 from 17 alleles).

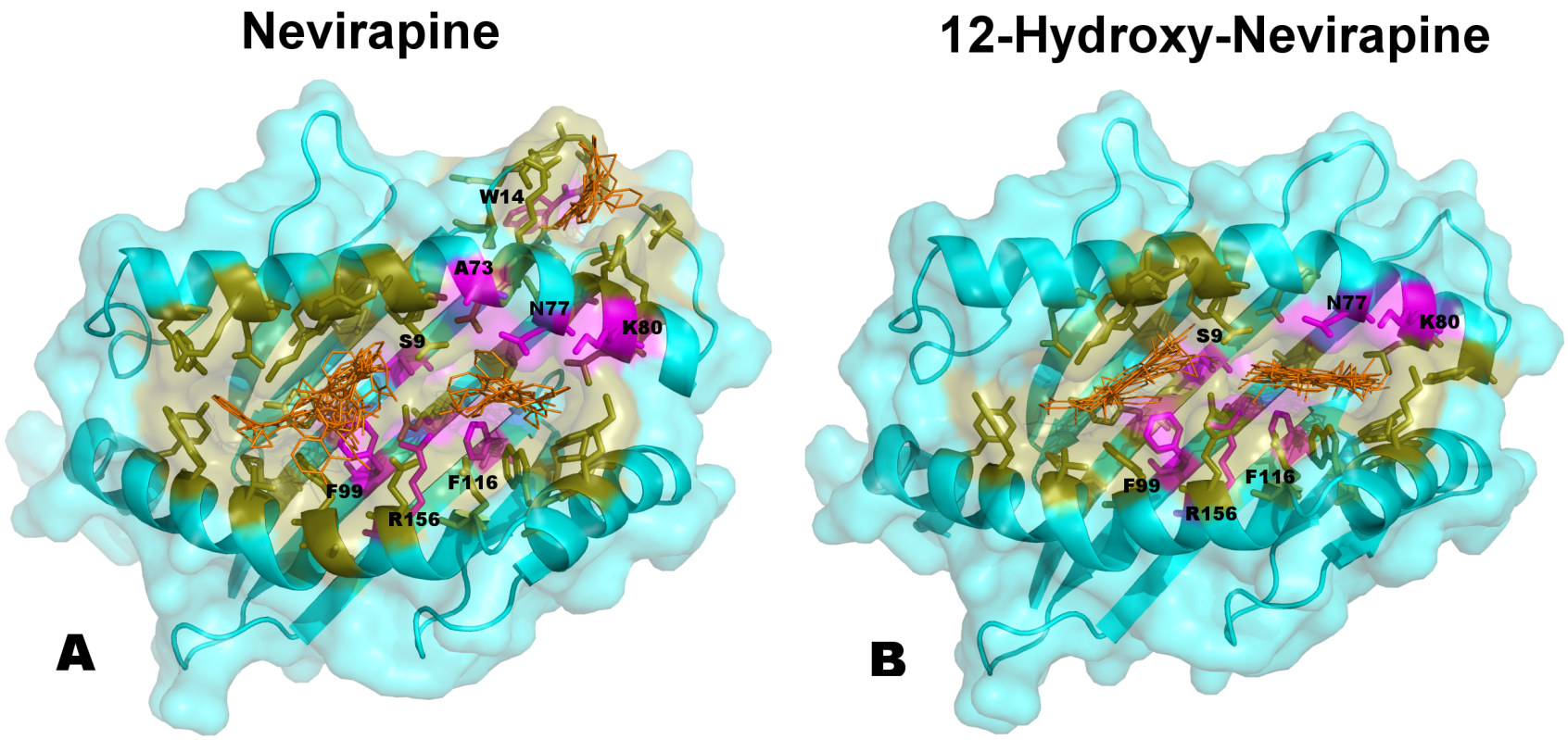


Figure 4.6: HLA-C*04:01 residues contacting all nevirapine (A) and 12-hydroxy-nevirapine (B) modes predicted by docking. Residues in magenta are polymorphic residues among the alleles present in Malawi cohort, for which more than 50% of the alleles are different than C*04:01, while residues in gold are all other residues contacting the predicted modes. This figure shows the same docking results as figures 4.4 and 4.5, but details all residues contacting the 20 conformations predicted for each ligand.

Table 4.4: Polymorphic residues among the alleles present in Malawi cohort for which more than 50% of the alleles are different than C*04:01 and their HLA carrier frequencies.

	Amino Acid Positions															
HLA Allele	**1**	**9**	**11**	**14[a]**	**49**	**73[a]**	**77**	**80**	**90**	**99**	**114**	**116**	**156**	**219**	**275**	**303**
C*04:01	G	S	S	W	E	A	N	K	D	F	N	F	R	W	K	M
C*02:10	C	Y	A	R	A	T	-	-	A	Y	D	S	W	R	E	V
C*03:02	-	Y	A	R	A	T	S	N	A	Y	D	S	L	-	E	V
C*03:03	-	Y	A	R	A	T	S	N	A	Y	D	Y	L	-	E	V
C*03:04	-	Y	A	R	A	T	S	N	A	Y	D	Y	L	-	E	V
C*06:02	C	D	A	R	A	-	-	-	-	Y	D	S	W	R	E	V
C*07:01	C	D	A	R	A	-	S	N	-	Y	D	S	L	R	E	V
C*07:02	C	D	A	R	A	-	S	N	-	S	D	S	L	R	E	V
C*07:04	C	D	A	R	A	-	S	N	-	Y	D	-	D	R	E	V
C*08:02	C	Y	A	R	A	T	S	N	A	Y	-	-	-	R	G	V
C*12:03	C	Y	A	R	A	-	S	N	A	Y	D	S	W	R	E	V
C*16:01	C	Y	A	R	A	T	S	N	A	Y	D	S	Q	R	E	V
C*17:01	-	Y	A	R	A	-	-	-	A	Y	-	-	L	R	-	V
C*18:01	C	D	A	R	A	-	-	-	-	-	-	-	-	-	-	V

'-' = Same as C*04:01 sequence. Columns in green represent residues contacting one or more conformation modes. [a] Positions 14 and 73 contact nevirapine, but not 12-hydroxy-nevirapine modes. Columns with a red outline are the top 3 residues containing more differences when comparing other alleles to C*04:01. For position 14 only C*04:07 has the same residue. For position 9, C*04:07 and C*14:02 have the same residue, and for position 99, C*04:07 and C*14:02 and C*18:01 have the same residue.

Thus, docking results have been capable to identify three possible binding sites for nevirapine, and two possible binding sites for 12-hydroxy-nevirapine. Each of those molecular sites has specific polymorphic positions with the potential to be key interactors with nevirapine in C*04:01 due to low replication of the amino acids in those positions in other alleles, especially the alleles found in the Malawi cohort where the nevirapine-induced ADR association with C*04:01 has been found. The next sections detail the configuration of those putative binding sites and the likely mechanisms leading to ADR.

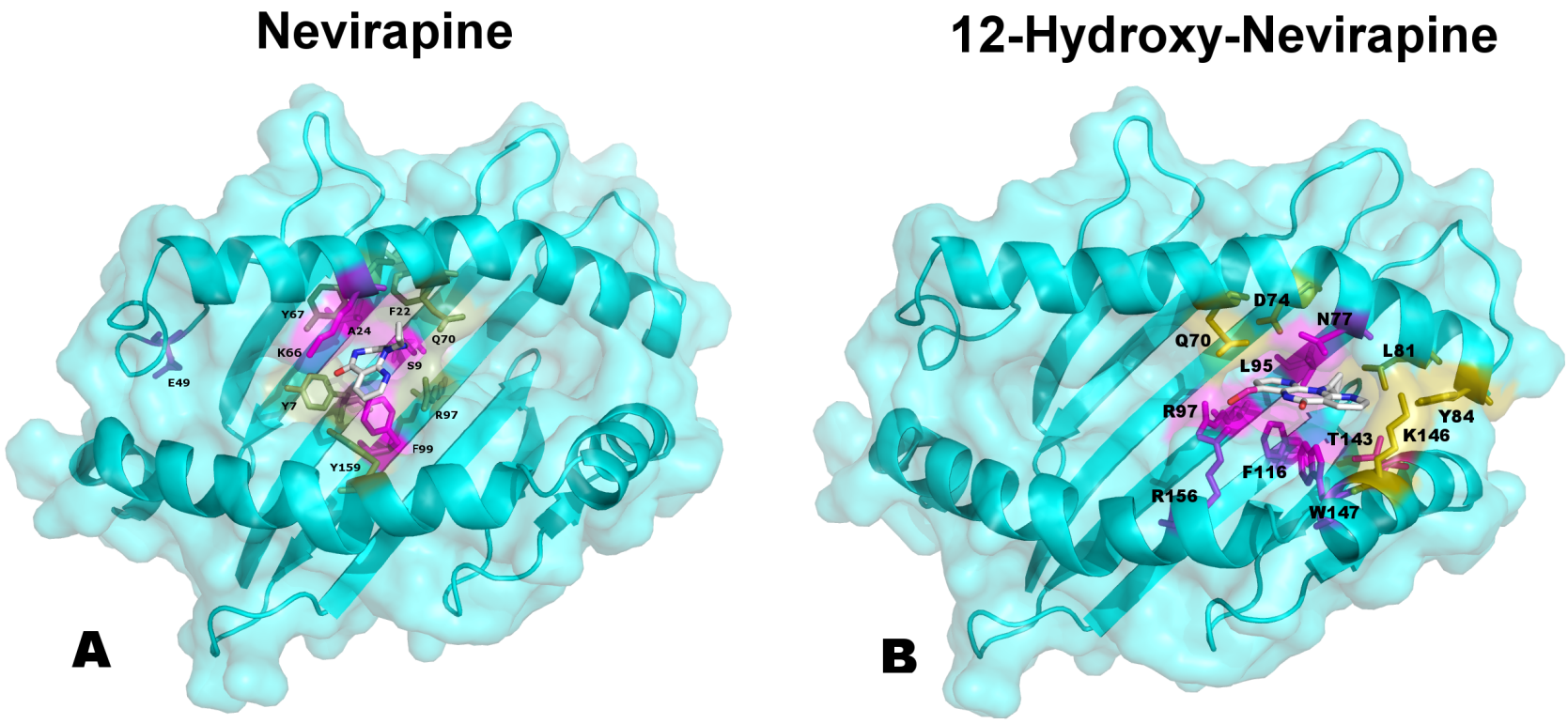


Figure 4.7: Nevirapine (A) and 12-hydroxy-nevirapine (B) first pose docked to HLA-C*04:01, highlighting the contacting residues. Residues in magenta are polymorphic residues among the alleles present in Malawi cohort, while residues in gold are all other residues contacting the predicted mode. Position 49 is the only residue shown in detail in nevirapine (A) that does not contact the predicted conformation.

4.4.1 Putative drug interaction with position 9 and 99 (NVP2)

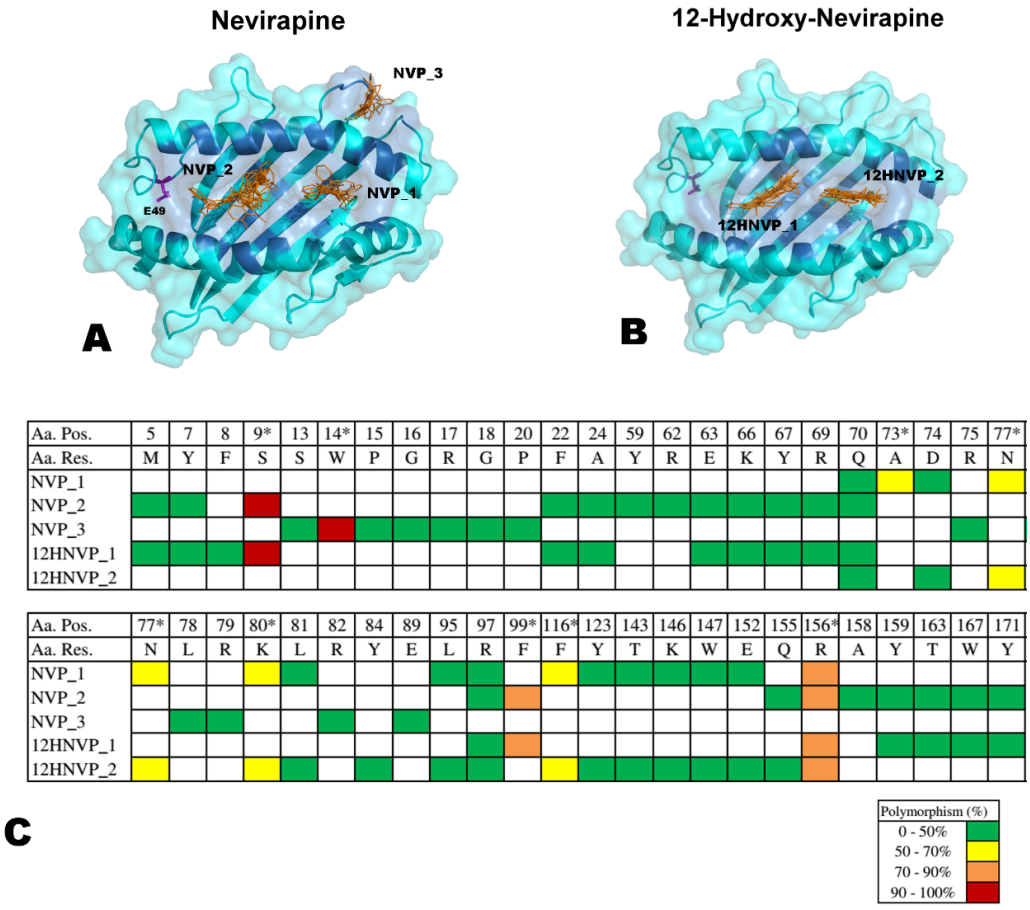
Docking results suggest a putative binding site that comprises residues S9 and F99 for both nevirapine and 12-hydroxy-nevirapine (Figure 4.8). For nevirapine docking, this binding site appears as the first conformation, i.e. the most likely conformation according to docking scoring system. For 12-hydroxy-nevirapine docking, it appears as the second conformation according to docking scores. From the 20 conformations generated for both ligands, 11 nevirapine conformations were predicted to bind this site in comparison to 9 12-hydroxy-nevirapine conformations.

Residues S9 and F99 have the potential to be key interactors with nevirapine or 12-hydroxy-nevirapine as they are strongly differentiated in other alleles present in the Malawi cohort in comparison to C*04:01 (Table 4.4). Excluding HLA-C*04:07, the only allele containing S9 is C*14:02 and F99 is present in C*14:02 and C*18:01. By analysing all HLA-C alleles which have been described so far it can be observed that the S9 and F99 in combination are only found in C*04 and C*14 alleles. This exclusivity of those residues to these two low resolution specificities suggests this molecular region as strong candidates for the actual binding site of both nevirapine and 12-hydroy-nevirapine, since C*14 and C*14:02 allele frequencies, both in world populations and restricting by sub-Saharan populations, are lower than C*04 (Figure 4.9), where larger sample sizes are required for finding significant associations.

Those two residues are also known constituent residues of some of the six pockets which are part of the HLA peptide binding groove. Residue S9 is part of two pockets, B and C, which accommodate peptide residues 2 and 6, respectively. Residue F99 is part of three pockets, being the residue with the higher participation in HLA pockets. It is part of pockets A, B and D, which accommodate peptide residues 1, 2 and 3 respectively [36].

Although there is little information regarding the mechanisms of HLA associations with ADRs, those residues have vital importance on the accommodation of the peptide in the peptide binding groove, determining their final conformation to be presented to T-cells. Recently, the mechanism of an association of B*57:01 and abacavir-induced ADR has been described, which is also linked to specific residues within the peptide binding groove. Asp114 and Ser116, where the first is part of pockets D and E, enable the binding of abacavir to the peptide binding groove, changing its conformation and chemical

properties, resulting in an alteration of the peptide repertoire binding this allele. Therefore, new endogenous peptides which were not previously identified as self-peptides during T-cell maturation start being exposed by B*57:01, leading to activation of circulating T-cells specific for those new peptides. Since S9 and F99 are also within the peptide-binding region and both are constituents of HLA pockets, a similar mechanism is likely to be the case in nevirapine scenario.



Aa. Pos.	5	7	8	9*	13	14*	15	16	17	18	20	22	24	59	62	63	66	67	69	70	73*	74	75	77*
Aa. Res.	M	Y	F	S	S	W	P	G	R	G	P	F	A	Y	R	E	K	Y	R	Q	A	D	R	N
NVP_1																								
NVP_2																								
NVP_3																								
12HNVP_1																								
12HNVP_2																								

Aa. Pos.	77*	78	79	80*	81	82	84	89	95	97	99*	116*	123	143	146	147	152	155	156*	158	159	163	167	171
Aa. Res.	N	L	R	K	L	R	Y	E	L	R	F	F	Y	T	K	W	E	Q	R	A	Y	T	W	Y
NVP_1																								
NVP_2																								
NVP_3																								
12HNVP_1																								
12HNVP_2																								

Figure 4.8: List of all contacting residues of predicted conformations of nevirapine and 12-hydroxy-nevirapine in C*04:01, separately for each ligand cluster referring to a putative binding site. (A) and (B) are references for the cluster names, and (C) details all residues contacting each cluster. * Polymorphic residues among the alleles present in Malawi cohort, for which more than 50% of the alleles are different than C*04:01. 'Aa. Pos.' = Amino acid position; 'Aa. Res.' = Amino acid residue; 'Polymorphism (%)' = Percentage of polymorphic differences in alleles found in Malawi cohort compared to C*04:01 (for this analysis, C*04:07 was excluded, since its only residue difference is position 49, which was shown in the present study unlikely to be the binding site for nevirapine or its metabolites). White (blank) squares are residues not contacting the predicted modes in each given cluster.

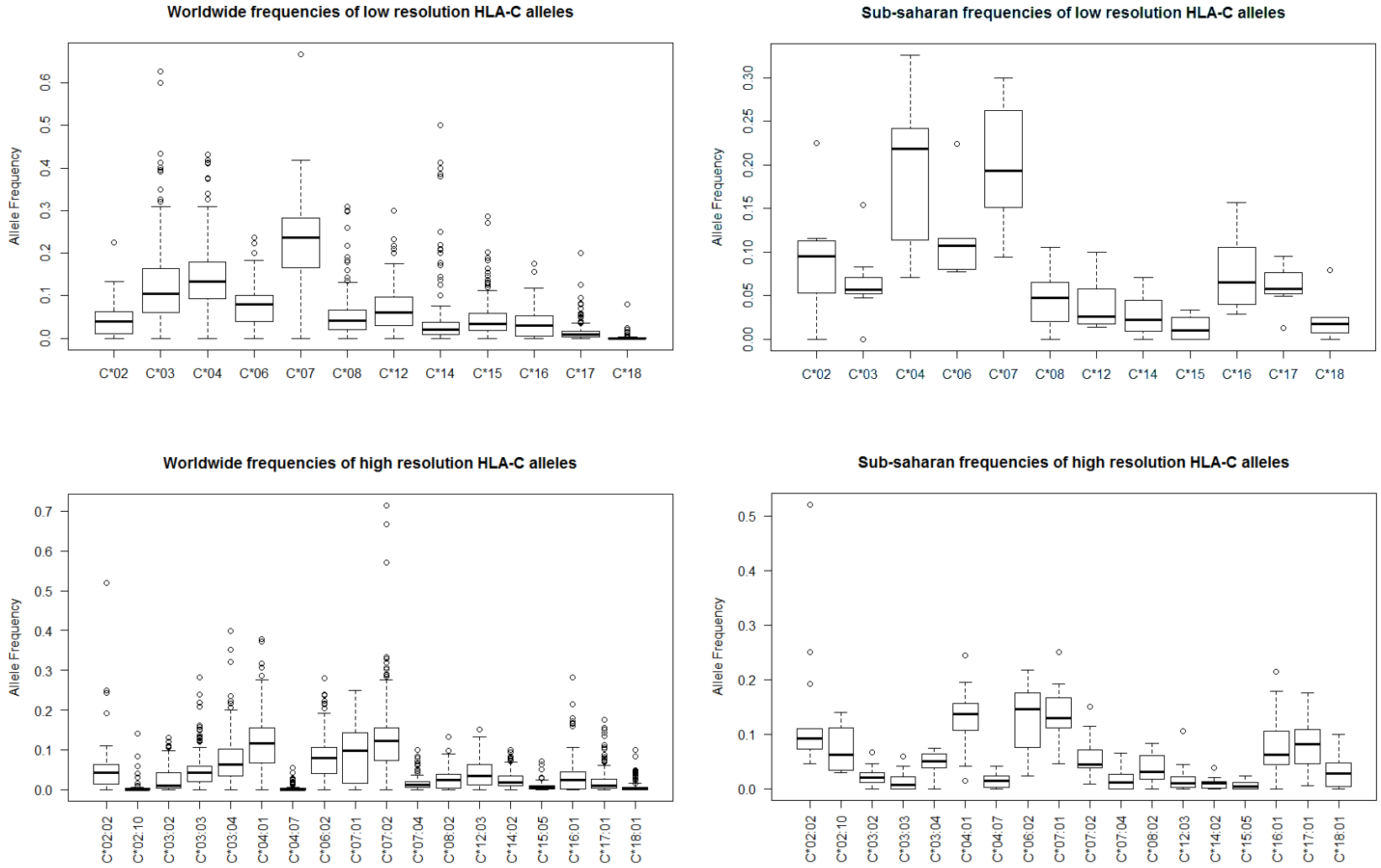


Figure 4.9: World distribution of allele frequencies of HLA-C alleles present in the Malawi cohort. Distributions are grouped by worldwide populations (left side) or restricted to sub-Saharan populations (right side) [58].

4.4.2 Putative drug interaction with position 14

Docking results show three regions in the C*04:01 where nevirapine can putatively bind (Figure 4.8). One of these regions is situated outside the peptide-binding region, but still on the top surface, close to residues within the peptide-binding region. The predicted nevirapine modes contact the residue Trp14, which is unique to C*04 (except for seven non-C*04 rare alleles). The fact that this region is also close to peptide-binding region, it is possible that the binding of nevirapine in this region could affect antigen presentation in different ways, by changing the chemical structure of the HLA-C molecule or the peptide resulting in different structures recognizable by T-cells as non-self, or by disturbing regions specific to NK cell recognition.

HLA-C*04:01, as other HLA-C alleles, are capable of interacting with KIR receptors, which are present on the surface of NK cells (Figure 4.10). This interaction results in signals that can be inhibitory or activating, depending on the KIR receptor type. NK cells will then be activated if there are enough activating signals to initiate a cytolytic response. HLA-C alleles are divided in C1 and C2 ligand groups. HLA-C1 ligands contain an asparagine in position 80, binding to KIR2DL2 and KIR2DL3 while HLA-C2 ligands contain lysine at position 80, binding to KIR2DL1 (C*04:01 allele is a C2 ligand). Most of the KIR receptors binding to HLA ligands are inhibitory receptors (including KIR2DL1), so when HLA class I molecules are downregulated or impaired, a response from NK cells is elicited due to the absence MHC class I interactions with inhibitory KIRs, and therefore absence of inhibitory signals. This 'missing-self' mechanism is and evolutionary development that provides an alternative defence against pathogens and tumours that reduce HLA class I molecule expression to avoid their antigen presentation to T-cells.

The main residue involved in HLA-C2 ligand interaction to inhibitory KIR2DL1 is Lys80 in HLA and Glu44 in KIR2DL1, but the molecular interface also includes other residues. In HLA-C*04:01, residue 14 is within a loop known to influence KIR interaction. Trp14 residue in HLA-C*04 has been determined as crucial for binding of KIR2DS4, by altering the conformation of the loop comprising residues 14 and 19, which is located close the KIR-HLA interaction site. Residue 19 is situated below Arg75, a residue that participates in the KIR2DL1 interaction with HLA-C2 ligands [175]. Although KIR2DS4 is an activating receptor, its interaction with C*04 alleles is weaker than KIR2DL1-C*04 interaction [176], thus the function of KIR2DS4 interaction has

not been yet clarified. The 14R>W substitution in HLA-C*04 may facilitate nevirapine binding on this site, resulting in an impairment of KIR2DL1 interaction, turning NK cells more prone to activate, resulting in a cytolytic response against cells containing HLA-nevirapine complexes. Both KIR2DL1 and KIR2DS4 have high worldwide frequencies [58]. NK and NKT cells are usually found in SJS/TEN blisters together with CD8+ T-cells. Despite it having been suggested that the inflammatory response in SJS/TEN is initiated by T-cells, it has also been argued that the quantity of T-cells found in those blisters is too low to generate the extensive inflammatory responses observed in this adverse drug reaction. Therefore, an immune response of NK cells due to its failure to recognize C*04 ligand through inhibitory KIR2DL1 as a result from the binding of nevirapine nearby Trp14 is possible, but it is unlikely to be the only mechanism leading to SJS/TEN.

4.4.3 Putative drug interaction with residue Phe116 and Arg156

Another predicted C*04:01 binding site for nevirapine and 12-hydroxi-nevirapine comprises residues Phe116 and Arg156 located within the peptide-binding region (Figure 4.8). In the Malawian cohort, both residues are also present in C*08:02 and C*18:01 alleles. C*08 has been previously associated with nevirapine-induced ADR in Asian and Caucasian populations, however none of those alleles were significantly associated with SJS/TEN in the Malawian cohort. C*08:02 was present in 13% of controls and 11% of SJS/TEN cases, while C*18:01 was present in 14% of controls and 8% of SJS/TEN cases. Despite there being no trend in the Malawian cohort for an association of C*08 and C*18 to SJS/TEN, C*08 and C*18 frequencies are considerably lower than C*04 in sub-Saharan populations (C*08 mean = 0.05, SD = 0.035; C*18 mean = 0.02, SD = 0.025; compared to C*04 mean = 0.19, SD = 0.088), and consequently C*04 frequency in the Malawian cohort is higher than other alleles, thus giving more support for discovery of disease associations. Furthermore, these positions have been previously described in one of the few studies detailing HLA binding sites for drugs causing ADRs. Residue Ser116 have been found to be crucial for binding of abacavir on HLA-B*57:01, while residue Trp156 has been related to binding of carbamazepine on HLA-B*15:02, both consisting of hypersensitivity mechanisms.

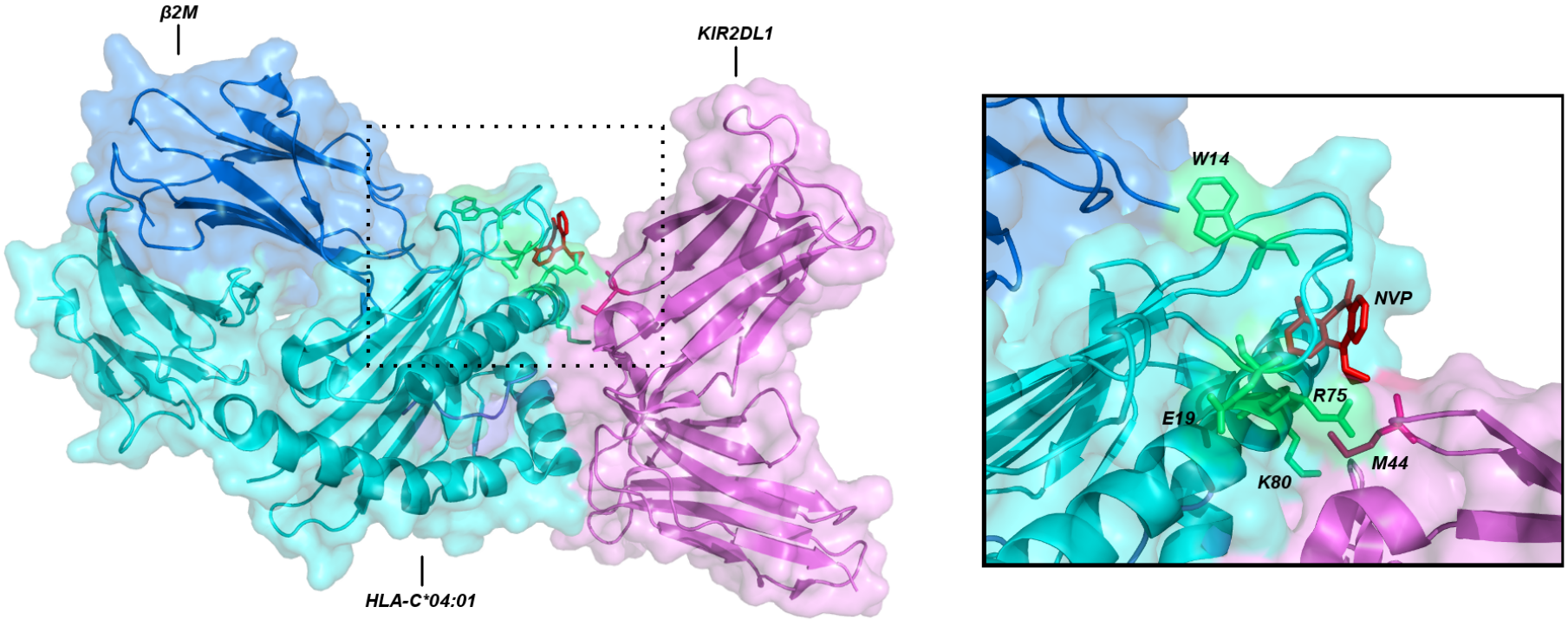


Figure 4.10: *In silico* nevirapine binding to the crystallographic structure of the interaction between HLA-C*04:01 and KIR2DL1, highlighting the residues in both molecules contacting nevirapine.

4.4.4 Critical evaluation of using docking tools for ADRs associated with HLA

The variability in the predicted binding sites and modes, in conjunction with their associated scores reported in this chapter, raises questions regarding the reliability of molecular docking tools and the significance of generated scores. Furthermore, since the PBG on the top of the HLA molecule comprised a cavity with optimal dimensions to fit small molecules such as drugs, one can assume that any molecule within a dimension range would be predicted to bind somewhere within the PBG, even if biologically they do not bind HLA molecules. Our group has subsequently followed up on these questions, in an attempt to optimise protocols and understand the pitfalls of docking [177].

Using drug-HLA associations with known binding sites as models, such as abacavir and carbamazepine, several molecular docking tools were benchmarked regarding their capability of predicting the binding site of a drug to their respective risk HLA alleles. The same analysis was also performed using control alleles with alternative residues in the binding sites to investigate if that changes the predicted binding sites. Results from those analyses showed that most docking tools were able to predict the correct binding site, and were also able to differentiate between risk and control alleles, where best scoring poses for control alleles were positioned further from the known binding site. However, several factors from variations in proteins generated by homology modelling, determination of flexible residues and adjustment of parameters were shown to have some influence on docking results.

Molecular docking tools should be used with caution, as they are not able to confirm that in reality a molecule binds their predicted binding sites. Nevertheless, they are certainly helpful tools to guide researchers regarding mechanisms involved in ADRs associated with HLA, and should be used in tandem with association studies and experimental evidence.

4.5 Conclusion

An association of a SNP encoding a Glu49 substitution in HLA-C*04:01 with nevirapine-induce hypersensitivity led to an initial consideration of its putative involvement in the ADR provoking mechanism, since it is nearly absent in non-C*04

alleles. While the GWAS study using a replication cohort added further weight to existing evidence of association of HLA-C*04:01 with nevirapine-induce hypersensitivity, molecular docking of nevirapine with this polymorphism shows no evidence of the involvement of this residue in nevirapine hypersensitivity. The docking analysis presented in this chapter, as well as analysis of HLA sequences and allele frequencies, has led to other HLA-C residues being identified as possible candidates to be influencing nevirapine interactions with HLA molecule for further investigations. It is acknowledged that there are some limitations to this work. First, it is known that docking can produce false positives, and the prediction of docking does not prove a given binding mechanism. Second, the work assumes that direct binding of nevirapine to the HLA molecule is the mechanism by which the ADR is induced. Given the strength of evidence supporting this mechanism with abacavir, this appears a reasonable assumption, however there is still controversy and discussion in the field. Despite these limitations, the work presented in this chapter has made steps towards suggesting strongly plausible hypotheses by which the ADR could be occurring, which can be followed up in new studies.

Chapter 5

Discussion, general conclusions and future work

The research described in this thesis is in the context of precision /personalised medicine, which is a growing field of clinical science and healthcare that incorporates comprehensive knowledge of inter-individual genetic, environmental and lifestyle variability to optimize clinical care. The distinctive complexity and clinical impact of immune genes and their related mechanisms makes them obvious targets for advances in this field.

Computational and informatics developments are central to integrate data needed to deliver tailored healthcare, and are essential for dealing with the continuous growth of diseases associations with immune genes, the development of better strategies for HLA matching for transplantation, and in uncovering immune mechanisms inducing HLA-linked hypersensitivity. The developments here described provide informatics resources for analysis of immunogenetic variability in diseases and transplantation and gives insights on how this variability shapes individual health. It also adds to the understanding of molecular mechanisms underlying nevirapine-induced hypersensitivity associated with HLA.

The development of a public database storing KIR and disease association studies (KIR and Diseases Database – KDDB) as part of a major immunogenetic database (Allele Frequency Net Database – AFND) integrates population gene frequency data with susceptibility to diseases, providing a tool for helping researchers in the field to better understand the role of those genes in modulating different disease types. A form of meta-analysis covering the curated data showed a relationship of KIR gene function and the type of disease investigated, where an enrichment of reported associations of activating KIR genes with autoimmune diseases, cancer and pregnancy complications was found. As in a more traditional meta-analysis, publication bias is potentially a limiting factor on this analysis.

Future work for the field of disease associations with immune genes should involve continuous gathering of KIR-disease associations and development of a database storing HLA and disease associations that could also be integrated with KDDB. For a better understanding of KIR interplay in influencing response to with diseases, further systematic analyses are needed. KIR and disease investigations would also benefit from further structural knowledge of the interaction of KIRs with ligands, with improved definitions of consequences of variable strength affinity in those interactions and influence of peptides presented by HLA molecules in the behaviour of those receptors.

HLA epitope frequencies database (EpFreq-DB) is another resource developed as part of AFND, which translates HLA allele data into HLA epitope data regarding population frequencies. With the increasing research towards structural approaches in transplantation matching making use of HLA epitopes, EpFreq-DB provides a tool for assessing the implementation of these methods in worldwide populations. Analysis of HLA epitope frequencies in several populations show that epitopes have distinct patterns of variability across populations, some being present in almost all individuals while others are almost never present. Epitopes showing higher population variability may be more relevant to be considered for donor-search strategies.

An explorative analysis of the HLA epitope matching was performed based on alloreactivity profiles from single-antigen bead (SAB) assays containing information regarding the presence of alloantibodies that can recognize HLA antigens present in those beads. Results from this analysis suggest that HLA antigens in SAB assays – comprising only a subset of existing polymorphisms – may not sufficiently cover variability in some populations. Since epitopes are shared across different alleles and loci, their application has the potential to improve granularity from SAB assay results. However, a better understanding of HLA epitopes is required, since most of the epitopes currently defined are originated from *in silico* analysis.

HLA polymorphisms have been also widely associated to hypersensitivity to several drugs. Recently, a strong association of HLA-C*04:01 with nevirapine-induced hypersensitivity has been reported. To investigate molecular aspects of this association, *in silico* docking analysis was performed for the interaction of nevirapine and a related metabolite with HLA-C*-04:01. Since a SNP substitution (Glu49) in C*04:01 was found associated with nevirapine-induced hypersensitivity, a mechanism role of this residue nearly unique to all C*04 alleles was initially hypothesized. Results from molecular

docking analysis showed no interaction of nevirapine with this position, but it identified other possible candidate residues to be involved in the nevirapine hypersensitivity mechanism.

Future generation of the crystallographic structure of nevirapine binding to C*04:01 is needed to verify those interactions. Additionally, following a rationale that nevirapine binds to residues within the HLA peptide-binding region and could be altering the peptide repertoire presented by C*04:01 molecules, further data of the peptides generated by C*04:01 and closely related alleles would also benefit the understanding of this mechanism.

Achievements from this thesis consist of steps towards a better understanding of mechanisms of immune variability in generating distinct outcomes in individual health, contributing to necessary knowledge regarding the predictive and preventive aspects for tailored healthcare.

Appendix A

This appendix outline the publications resulting from this work in the next pages.

DATABASE
The Journal of Biological Databases and Curation

Database, Vol. 2013, Article ID bat021, doi:10.1093/database/bat021

Original article

A database for curating the associations between killer cell immunoglobulin-like receptors and diseases in worldwide populations

Louise Y. C. Takeshita[1,*], Faviel F. Gonzalez-Galarza[1], Eduardo J. M. dos Santos[2], Maria Helena T. Maia[2], Mushome M. Rahman[1], Syed M. S. Zain[1], Derek Middleton[3] and Andrew R. Jones[1]

[1]Institute of Integrative Biology, Functional and Comparative Genomics, University of Liverpool, Liverpool, L69 7ZB, UK, [2]Human and Medical Genetics, Federal University of Pará, Belém-Pa, 66075-110, Brazil and [3]Transplantation Immunology, Royal Liverpool and Broadgreen University Trust and University of Liverpool, Liverpool, L7 8XP, UK

***Corresponding author:** Tel: +44 151 795 4555; Fax: +44 151 795 4410; Email: L.Takeshita@liverpool.ac.uk

Submitted 30 November 2012; Revised 28 January 2013; Accepted 12 March 2013

Citation details: Takeshita,L.Y.C., Gonzalez-Galarza,F.F., Santos,E.J.M., *et al*. A database for curating the associations between killer cell immunoglobulin-like receptors and diseases in worldwide populations. *Database* (2013) Vol. 2013: article ID bat022; doi:10.1093/database/bat022.

The killer cell immunoglobulin-like receptors (KIR) play a fundamental role in the innate immune system, through their interactions with human leucocyte antigen (HLA) molecules, leading to the modulation of activity in natural killer (NK) cells, mainly related to killing pathogen-infected cells. KIR genes are hugely polymorphic both in the number of genes an individual carries and in the number of alleles identified. We have previously developed the Allele Frequency Net Database (AFND, http://www.allelefrequencies.net), which captures worldwide frequencies of alleles, genes and haplotypes for several immune genes, including KIR genes, in healthy populations, covering >4 million individuals. Here, we report the creation of a new database within AFND, named KIR and Diseases Database (KDDB), capturing a large quantity of data derived from publications in which KIR genes, alleles, genotypes and/or haplotypes have been associated with infectious diseases (e.g. hepatitis C, HIV, malaria), autoimmune disorders (e.g. type I diabetes, rheumatoid arthritis), cancer and pregnancy-related complications. KDDB has been created through an extensive manual curation effort, extracting data on more than a thousand KIR-disease records, comprising >50 000 individuals. KDDB thus provides a new community resource for understanding not only how KIR genes are associated with disease, but also, by working in tandem with the large data sets already present in AFND, where particular genes, genotypes or haplotypes are present in worldwide populations or different ethnic groups. We anticipate that KDDB will be an important resource for researchers working in immunogenetics.

Database URL: http://www.allelefrequencies.net/diseases/

Introduction

Natural killer (NK) cells are bone marrow-derived lymphocytes that play an active role in the innate immune system by interacting with human leucocyte antigen (HLA) class I molecules to kill pathogen-infected cells (1). Initially, NK cells were discovered as a result of their ability to target and kill tumour cell lines that expressed little or no HLA class I molecules (2). It is now known that the killing function in NK cells is dependent on a mixture of activating and

inhibitory receptors present on the membrane and the interaction with their HLA ligand (3). Two main types of receptors are found in NK cells, C-type lectin-like (NKG2D, CD94/NKG2C, CD94/NKG2A) and the immunoglobulin-like superfamily (KIR, CD16, NKp30, NKp44, etc.). In the latter, the killer cell immunoglobulin-like receptors (KIR) that mostly bind Major histocompatibility complex (MHC) class I molecules have been shown to be the most polymorphic. Despite most of the NK cell receptors binding MHC class I-related molecules, several Ig-like receptors bind non-HLA ligands, for example, CD16 binds IgG, triggering an activating response, and NKp44, NKp30 and NKp46 are activating receptors that bind molecules expressed by pathogens and self-ligands (4–8).

The KIR gene cluster is located in the leucocyte receptor complex (LRC) at position 19q13.4 (4, 5). To date, 16 *KIR* genes have been identified, coding for receptors with activating (*KIR2DS1*, *KIR2DS2*, *KIR2DS3*, *KIR2DS4*, *KIR2DS5A/B* and *KIR3DS1*) or inhibitory (*KIR2DL1*, *KIR2DL2*, *KIR2DL3*, *KIR2DL5A*, *KIR2DL5B*, *KIR3DL1*, *KIR3DL2* and *KIR3DL3*) function, with *KIR2DL4* appearing to have both functions. Two pseudogenes *KIR2DP1* and *KIR3DP1* have also been identified (9). Structurally, the activating and inhibitory functions of KIR are related to the length of their cytoplasmic tail that can be short (S) or long (L), distinguished in the nomenclature (9).

Variation in KIR can result from a different gene and/or allele content of an individual (10), giving rise to haplotype diversity and leading to a very large number of different genotypes that have been observed (presence/absence of KIR genes). The KIR genes *KIR2DL4*, *KIR3DL2*, *KIR3DL3* and *KIR3DP1* are present in nearly all individuals with a few exceptions (11), and are commonly known as 'framework' genes. The frequencies of inhibitory and activating genes vary in different populations, as reviewed in (11). A 24-kb band using HindIII digestion and Southern blot analysis distinguishes the haplotypes, termed A and B, that make up the genotype (12). The A haplotype is generally non-variable in its gene content—framework genes plus *KIR2DL1*, *KIR2DL3*, *KIR2DS4* and *KIR3DL1*—although occasionally one of these genes may be missing (11). In contrast, the B haplotype contains one or more of the genes encoding activating KIRs—*KIR2DS1/2/3/5* and *KIR3DS1*—and the genes encoding inhibitory KIRs—*KIR2DL5A/B* and *KIR2DL2*. In B haplotypes, variability is created by both the presence/absence of a gene and by allelic variation; in contrast, A haplotypes owe much of their variability to allele content (11). At the last release of IPD-KIR (Release 2.4.0), there were 601 KIR alleles reported (13). B haplotypes tend to be more prevalent in non-Caucasian populations, such as Australian Aborigines and Asian Indians, whereas in Caucasian populations, ~55% will have one and 30% two A haplotypes (14, 15). It is thought that populations with higher frequencies of B haplotypes are those under strong pressure from infectious diseases. Such extensive diversity among modern populations may indicate that geographically distinct diseases have exerted recent or perhaps on-going selection on KIR repertoires. From a practical viewpoint, this makes the choice of controls very important for all disease association studies.

To collect allele, haplotype and genotype frequencies of several immune genes in different healthy human populations, the Allele Frequency Net Database (AFND) was developed (16). AFND stores large sets of data regarding HLA, KIR major histocompatibility complex class I chain related (MIC) and cytokine gene polymorphisms, and has shown to be frequently used in the immunogenetics field, receiving 200 hits per day on average. To date, 398 different KIR genotypes in 12 856 individuals from 109 populations have been reported to AFND.

Owing to its high level of polymorphism, many infectious and autoimmune diseases have been associated with KIR genes in different ways, e.g. associations with single genes (or single alleles) to associations with groups of genes and full genotypes (17–20). A disease association is defined as a statistically significant association between a genetic element (gene, allele, genotype, etc.) with a given disease outcome, either positive or negative i.e. the genetic profile makes the disease more likely/severe or less likely/severe than the control population. As such, the development of a database to store data regarding disease associations with those genes is a necessary step towards a more effective comprehension of such complex data. As KIR disease association studies are in its infancy compared with HLA, a decision was made to start collecting KIR disease associations, as a new module within AFND.

Materials and methods

Data curation

The first step towards creation of KDDB was the collection and extraction of data from peer-reviewed publications, following the workflow shown in Figure 1. Published KIR and disease association studies were extracted from the HuGE Navigator (version 2.0) (21), which is a web-based tool enabling searches of the scientific literature for studies on genetic associations with diseases. The HuGE Navigator makes use of the MeSH (Medical Subject Headings) terminology, which contains standardized keywords associated with clinically related published studies. In KDDB, we loaded MeSH terms that describe specific diseases with which associations have been found. Manual curation was performed to extract relevant data from retrieved studies. A set of consistent rules were applied to ensure that different curators extracted data in the same way (Figure 1). All studies identified based on the relevant MeSH terms were analysed and inserted into KDDB unless they did not pass

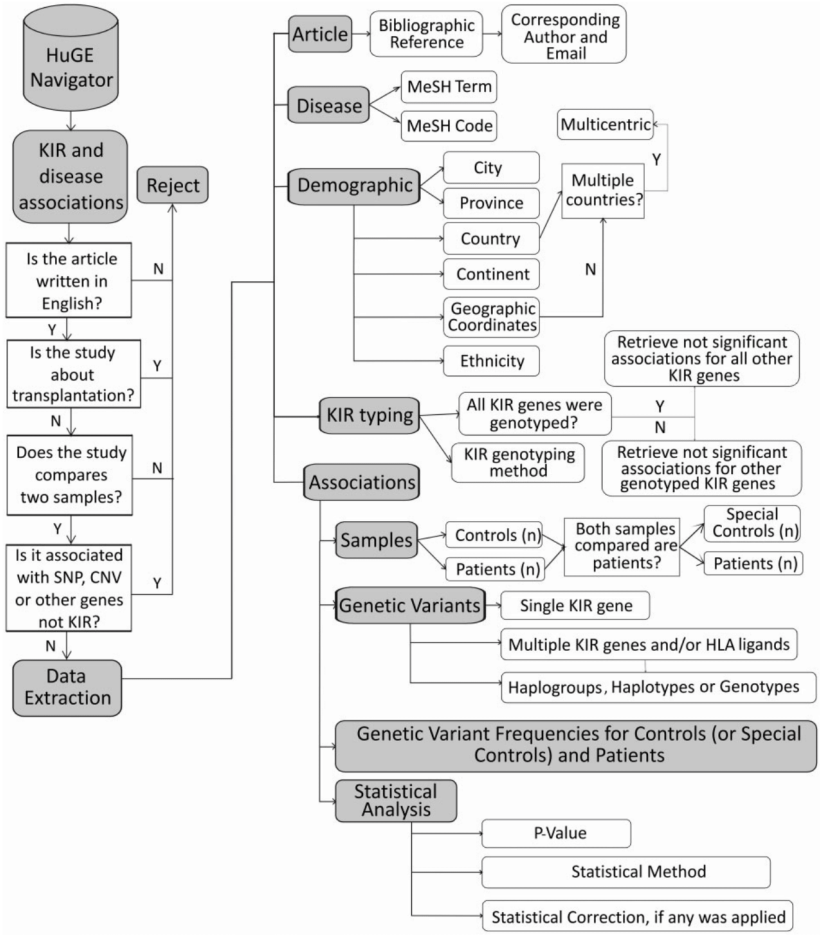


Figure 1. The data curation pipeline, the types of data that were extracted from each publication and the submission workflow developed within KDDB.

one of the following criteria (as also shown on Figure 1): (i) the article was not written in English, as we do not have the capability to translate articles at present, (ii) the study design was not based on a gene frequency comparison between two samples with different clinical outcomes (future updates to KDDB will attempt to include more complex study designs), (iii) the article identified by the HuGE Navigator was not in fact related to KIR (i.e. misidentified), (iv) the study was not related to a disease specifically, but instead describe transplantation outcomes. Studies associating transplantation outcome and KIR have heterogeneous designs—some studies associate KIR–ligand

Database, Vol. 2013, Article ID bat021, doi:10.1093/database/bat021

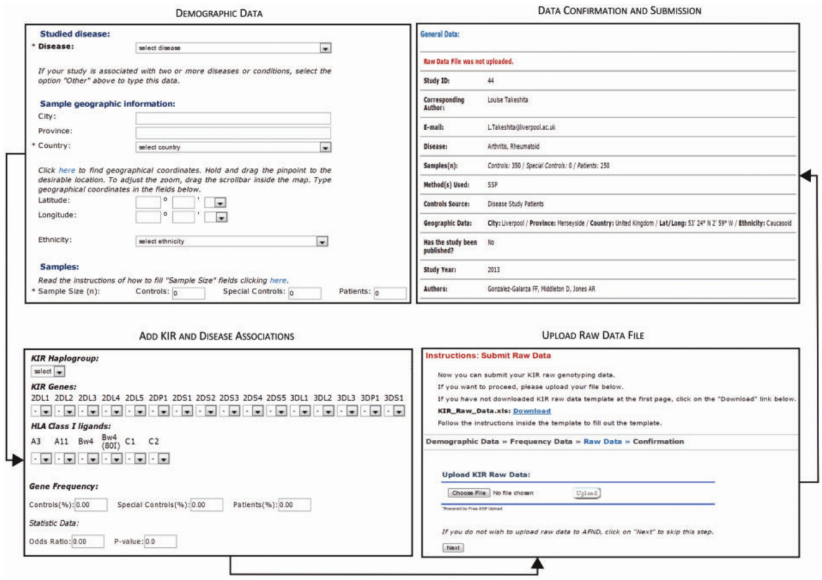


Figure 2. Screenshots of the data submission pipeline within KDDB.

matches/mismatches based on recipients and donors samples, and others correlate the risk of relapse with KIR combinations. These study designs are under evaluation to ascertain whether they can either fit the existing KDDB schema, or will be stored in a different database schema in the future. A data validation pipeline was created to ensure that quantitative data and metadata had been correctly extracted from each publication, involving two curators reviewing the same source publication to reduce the chance for misinterpretation or copy–paste errors.

Implementation

The back end of the database was developed using a relational database schema. Users can connect to the database using the most common web browsers. The web interface of KDDB has been created to allow users to retrieve or query the database and to submit new data sets. For that purpose, interactive web pages for querying and submitting data were developed using the Active Server Pages (ASP) scripting environment and JavaScript language. The graphical display was designed using HyperText Markup Language (HTML) and Cascading Style Sheets (CSS), ensuring that the page will be viewable in most used web browsers. The data submission pipeline will be an important feature for future updates to KDDB, as we recognize the benefits of obtaining community input, including unpublished data sets (see Discussion).

To submit studies to KDDB, a submission form pipeline was developed, which can be accessed through the AFND homepage by the menu 'Submissions' and the submenu 'Add KIR and disease association study' (Figure 2). This web form consists of four steps. The first step captures summary information about the study including the number of patients and controls. Information is also captured on the geographic location of the population, the ethnicity and the bibliographic reference. The second step captures the disease association data—the genes, alleles, haplotypes, KIR-HLA ligands, etc., the disease name, the frequency of patients and controls exhibiting the given genetic profile and the results of the statistical test. The third step (optional) allows users to upload anonymized raw data (the KIR genetic profile of every individual in the study). The fourth step allows users to review their data and submit. We anticipate that the submission pipeline will become an important tool for users to submit their own unpublished studies or studies missed in the curation process.

Table 1. Summary of data stored in the KDDB

Disease type	No. of studies	No. of records (T/S)	No. of individuals
Infectious	36	145/145	15 813
Autoimmune or Idiopathic	61	589/274	30 888
Neoplasias	9	135/47	5791
Pregnancy related	11	167/39	4879
Total	113[a]	1027/496[a]	56 214[a]

[a]Some of the studies fall into more than one disease type category, e.g. tumours originated from viral infections.
T, total records; S, significant associations.

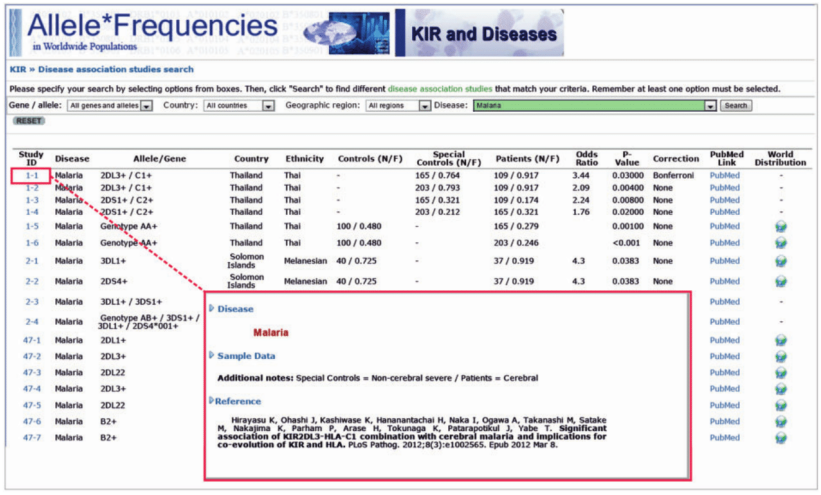


Figure 3. The query interface within KDDB, showing the additional detail about a given association study retrieved by following the hyperlink.

Results

Website organization and content

Using the HuGE Literature Finder tool, 159 articles remained after applying the exclusion criteria detailed in Figure 1. From all the articles, a total of 1027 KIR–disease associations were captured from 113 articles. A set of 46 articles was removed at this stage owing to studies lacking mandatory data/metadata or the numerical data were inaccessible, for example displayed only on charts. The genetic associations identified in this data compilation included those with single KIR genes, profiles of combined KIR genes and / or HLA class I ligands, and full KIR genotypes. In total, 70 unique MeSH terms have been associated with KIR across the studies in the present database. Classifying the studies by the main disease groups, 36 studies are related to infectious diseases, 61 studies are related to autoimmune or idiopathic diseases, 11 studies are related to pregnancy and 9 articles are related to cancer (Table 1). From these studies, a total of 1027 KIR records were inserted into KDDB, of which 496 are statistically significant KIR–disease associations.

The KIR and Diseases Database is part of Allele Frequencies Net Database, and can be accessed through Allele Frequencies Net homepage (http://www.allelefrequencies.net/) using the menu 'KIR' and the submenu 'KIR and disease associations' or via a direct URL access at http://www.allelefrequencies.net/diseases/. The website interface allows the user to retrieve and query KIR and disease associations applying a collection of filters. The user can restrict the search by gene or allele, country of origin of studied samples, continent of origin of studied samples or studied disease. Those filters can be applied alone or used in combination.

Results from a query are retrieved in a table format, with each row being a different disease association with KIR (Figure 3). In each row, the following information is displayed: (i) row number, (ii) the associated MeSH term, (iii) the country of origin of the sample, (iv) the associated KIR profile, (v) the sample size and gene frequencies for controls and patients, (vi) odds ratio value, (vii) *P*-value and (viii) statistical method used in comparisons. A link is provided, by clicking on the population name, to show the demographic information on the disease and corresponding control populations. As for normal populations in AFND, individual KIR gene frequencies or haplotype frequencies can be plotted on world maps. This enables a user to interpret disease association risks for KIR profiles in a geographic, ethnic group or individual population-based context. Additional functionality is under development for linking to external resources including to the IPD-KIR database (www.ebi.ac.uk/ipd/kir/), where the sequences and official nomenclature are maintained.

Discussion

In our original search for frequency data in AFND in normal populations, we sourced publication data from >65 peer-reviewed journals—a complete list of data sets and journals may be consulted at http://www.allelefrequencies.net/datasets.asp. However, many disease studies, especially those that do not find statistically significant associations, are not published, and there is a risk that resources such as KDDB could suffer from publication bias. As such, we are contacting colleagues working in this field with a request to provide their data, even if it is unpublished or does not contain a statistically significant association. As unpublished studies are added to KDDB, we will add a filter to the query page allowing users to exclude these data sets if they wish to ensure quality control. We are also requesting users to upload anonymized raw data (individual KIR type and HLA ligands) to enable improved quality control measures (such as validation of frequency calculations) and to enable advanced analyses of the data. For example, having the individual data available will allow analyses such as looking at disease associations in the centromeric or the telomeric regions. It is known there is extensive linkage disequilibrium between KIR genes, but this exists separately in the centromeric half and the telomeric half (22). There is little linkage disequilibrium between the two halves, and the genes KIR3DP1 and KIR2DL4 are at the division between centromeric and telomeric sections.

We already have some associations in KDDB derived from the presence of the KIR gene and its HLA ligand, and it will be important to expand this collection and include raw data. Studies have shown that although KIR and HLA genes are coded on different chromosomes, there are correlations (both negative and positive) between the presence of the KIR gene and corresponding presence/absence of the ligand (23, 24). These correlations have been shown to be important in diseases. For example, a reciprocal relationship exists in populations between the frequencies of the KIR A haplotype and the HLA-C2 group. This is believed to be due to an increased risk of pre-eclampsia when the mother lacks the AA haplotype and the foetus carried the HLA-C2 group (25). Further, KIR2DL3 was found to be associated with the development of cerebral malaria when the HLA-C1 ligand is present (26).

The first release of KDDB reported here includes only data we have extracted and curated from the scientific literature, identified by the HuGE Navigator. We are aware that the HuGE Navigator does not retrieve all studies and as such we are using other search strategies, for example via Pubmed and Web of Knowledge to locate studies missed in the first pass curation process. We have currently excluded studies that do not fit into the simple model of a case–control disease association study. Capturing more complex stratification studies is possible in KDDB, but will necessitate either some loss of granularity of the data, or the development of a much more complex schema and display interface. KDDB also does not yet contain any raw data, although the schema and submission pipeline are developed and tested to receive such data. KDDB is going to be maintained through our own data mining and curation efforts and through the submission of data from contributing laboratories (with suitable quality control procedures, as currently used in AFND). We are also exploring holding community workshops in the future to collect and collate data sets not yet in the public domain.

At present, we are not aware of any other site designed for public deposition of the raw data associated with immunogenetic disease association studies, and thus, these are not available for public analysis. The release of KDDB provides a new home for this raw data, and we encourage research groups that have published studies in the past, or those in the process of publishing new studies, to deposit the raw data within KDDB. We also encourage feedback from the scientific community on the utility of the data submission and query interface and the general approach we have taken to curation.

Conclusions

Over the last 10 years of existence, AFND has provided the immunogenetics and histocompatibility community with an online repository for the examination of frequencies in different healthy populations. With the development of the KDDB, our aim is to cover disease studies that have been associated with KIR genes and to include studies in which no significant association has been found, to avoid publication bias. In the future, we will extend the alleles covered to include other loci and new data sets as they are published. We anticipate that KDDB will greatly facilitate meta-analyses and data re-use to understand the underlying function of KIR genes in a variety of disease processes.

Funding

PhD studentship from CNPq (National Council for Scientific and Technological Development—Brazil) (to L.Y.C.T.). Funding for open access charge: University of Liverpool Library.

Conflict of interest. None declared.

References

1. Caligiuri,M.A. (2008) Human natural killer cells. *Blood*, **112**, 461–469.
2. Ljunggren,H.G. and Karre,K. (1990) In search of the 'missing self': MHC molecules and NK cell recognition. *Immunol. Today*, **11**, 237–244.
3. Moretta,L., Biassoni,R., Bottino,C. *et al*. (2000) Human NK-cell receptors. *Immunol. Today*, **21**, 420–422.
4. Liu,W.R., Kim,J., Nwankwo,C. *et al*. (2000) Genomic organization of the human leukocyte immunoglobulin-like receptors within the leukocyte receptor complex on Chromosome 19q13.4. *Immunogenetics*, **51**, 659–669.
5. Wende,H., Colonna,M., Ziegler,A. *et al*. (1999) Organization of the leukocyte receptor cluster (LRC) on human Chromosome 19q13.4. *Mamm. Genome*, **10**, 154–160.
6. Bashirova,A.A., Martin,M.P., McVicar,D.W. *et al*. (2006) The killer immunoglobulin-like receptor gene cluster: tuning the genome for defense. *Annu. Rev. Genomics Hum. Genet.*, **7**, 277–300.
7. Brusilovsky,M., Rosental,B., Shemesh,A. *et al*. (2012) Human NK cell recognition of target cells in the prism of natural cytotoxicity receptors and their ligands. *J. Immunotoxicol.*, **9**, 267–274.
8. Moretta,A., Marcenaro,E., Parolini,S. *et al*. (2008) NK cells at the interface between innate and adaptive immunity. *Cell Death Differ.*, **15**, 226–233.
9. Marsh,S.G., Parham,P., Dupont,B. *et al*. (2003) Killer-cell immunoglobulin-like receptor (KIR) nomenclature report, 2002. *Immunogenetics*, **55**, 220–226.
10. Wilson,M.J., Torkar,M., Haude,A. *et al*. (2000) Plasticity in the organization and sequences of human KIR/ILT gene families. *Proc. Natl. Acad. Sci.*, **97**, 4778–4783.
11. Middleton,D. and Gonzelez,F. (2010) The extensive polymorphism of KIR genes. *Immunology*, **129**, 8–19.
12. Uhrberg,M., Valiante,N.M., Shum,B.P. *et al*. (1997) Human diversity in killer cell inhibitory receptor genes. *Immunity*, **7**, 753–763.
13. Robinson,J., Halliwell,J.A., McWilliam,H. *et al*. (2013) IPD—the Immuno Polymorphism Database. *Nucleic Acids Res.*, **41**, D1234–D1240.
14. Rajalingam,R., Krausa,P., Shilling,H.G. *et al*. (2002) Distinctive KIR and HLA diversity in a panel of north Indian Hindus. *Immunogenetics*, **53**, 1009–1019.
15. Norman,P.J., Carrington,C.V., Byng,M. *et al*. (2002) Natural killer cell immunoglobulin-like receptor (KIR) locus profiles in African and South Asian populations. *Genes Immun*, **3**, 86–95.
16. Gonzalez-Galarza,F.F., Christmas,S., Middleton,D. *et al*. (2011) Allele frequency net: a database and online repository for immune gene frequencies in worldwide populations. *Nucleic Acids Res.*, **39**, D913–D919.
17. Jamil,K.M. and Khakoo,S.I. (2011) KIR/HLA interactions and pathogen immunity. *J. Biomed. Biotechnol.*, **2011**, 298348.
18. Khakoo,S.I. and Carrington,M. (2006) KIR and disease: a model system or system of models? *Immunol. Rev.*, **214**, 186–201.
19. Blackwell,J.M., Jamieson,S.E. and Burgner,D. (2009) HLA and infectious diseases. *Clin. Microbiol. Rev.*, **22**, 370–385.
20. Kulkarni,S., Martin,M.P. and Carrington,M. (2008) The Yin and Yang of HLA and KIR in human disease. *Semin. Immunol.*, **20**, 343–352.
21. Yu,W., Gwinn,M., Clyne,M. *et al*. (2008) A navigator for human genome epidemiology. *Nat. Genet.*, **40**, 124–125.
22. Gourraud,P.A., Meenagh,A., Cambon-Thomsen,A. *et al*. (2010) Linkage disequilibrium organization of the human KIR superlocus: implications for KIR data analyses. *Immunogenetics*, **62**, 729–740.
23. Parham,P., Norman,P.J., Abi-Rached,L. *et al*. (2012) Human-specific evolution of killer cell immunoglobulin-like receptor recognition of major histocompatibility complex class I molecules. *Philos. Trans. R Soc. Lond. B Biol. Sci.*, **367**, 800–811.
24. Guinan,K.J., Cunningham,R.T., Meenagh,A. *et al*. (2010) Signatures of natural selection and coevolution between killer cell immunoglobulin-like receptors (KIR) and HLA class I genes. *Genes Immun.*, **11**, 467–478.
25. Hiby,S.E., Apps,R., Sharkey,A.M. *et al*. (2010) Maternal activating KIRs protect against human reproductive failure mediated by fetal HLA-C2. *J. Clin. Invest.*, **120**, 4102–4110.
26. Hirayasu,K., Ohashi,J., Kashiwase,K. *et al*. (2012) Significant association of KIR2DL3-HLA-C1 combination with cerebral malaria and implications for co-evolution of KIR and HLA. *PLoS Pathog*, **8**, e1002565.

KIR and Disease Associations in the Allele Frequencies Net Database

www.allelefrequencies.net/diseases

*Louise Y C Takeshita, MSc1, Faviel F Gonzalez-Galarza, PhD1, Andrew R Jones, PhD1 and Derek Middleton, PhD, DSc, FRCPath2,**

1Institute of Integrative Biology, University of Liverpool, Liverpool, United Kingdom; 2Transplantation Immunology, Royal Liverpool and Broadgreen University Trust and University of Liverpool, Liverpool, United Kingdom.

** Corresponding author: Derek.Middleton@rlbuht.nhs.uk, +44 151 924 0655*

The killer cell-immunoglobulin-like receptors (KIR) play a fundamental role in the innate immune system, through their interactions with human leukocyte antigen (HLA) molecules. The interactions lead to the modulation of activity in natural killer (NK) cells, mainly related to killing pathogen infected cells. KIR molecules are expressed from a gene cluster located in the leukocyte receptor complex (LRC) at position 19q13.4, which shows extensive polymorphism, varying both in the number of genes an individual carries and in the number of alleles identified. To date, 15 KIR genes have been identified coding for activating and/or inhibitory receptors [1, 2].

Due to their function within the immune system, KIR genes have been associated with a large variety of diseases, and the number of publications reporting associations per year is continually increasing (Figure 1). However, the complex organization of the KIR gene cluster, its variability and its interaction with HLA class I molecules leads to a large number of possible genetic configurations that could be associated with a particular disease. As a result, most KIR and disease studies report sets of associations, which can also be combined with different disease classifications and symptoms. The complexity of the association data, and the increasing number of publications, was the main motivation towards the development of a database to store relevant data from KIR and disease studies.

At present, interpretation of worldwide immunogenetic variation, can be performed via the Allele Frequency Net Database (AFND, http://www.allelefrequencies.net), which captures worldwide frequencies of alleles, genes and haplotypes for several immune genes, including HLA and KIR, in healthy populations [3]. As a next step in helping to understand disease modulation by immune genes, a new database within AFND has been developed, named KIR and Diseases Database (KDDB). KDDB captures data derived from publications in which KIR genes, alleles, genotypes and/or haplotypes have been associated with infectious diseases (e.g. hepatitis C, HIV, malaria), autoimmune disorders (e.g. type I diabetes, rheumatoid arthritis), cancer and pregnancy-related complications. KDDB has been created through an extensive manual curation effort, extracting data on more than a thousand *KIR*-disease associations, comprising more than 50,000 individuals [4].

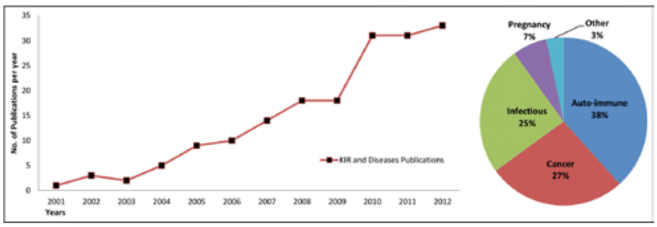


Figure 1: The graph shows an increasing trend in the number of publications of KIR and disease association studies. The pie chart summarizes disease MeSH (Medical Subject Headings) terms [5] which have been associated with KIR genes, from a total of sixty terms. Source data is from KIR and Diseases Database (KDDB) [4].

The primary source for publications was the HuGE Navigator, which is a web-based tool enabling searches of the scientific literature for studies on genetic associations with diseases [6]. From the output of the HuGE Navigator we retrieved 159 primary research articles on KIR and disease, which were manually curated to produce 1027 KIR disease-associations records. The following data types were extracted from articles and can be retrieved in KDDB: (i) disease name, (ii) geographic region, (iii) ethnicity, (iv) KIR genes, alleles and/or HLA ligands, (v) sample size and frequencies, (vi) statistical results (P-value, odds ratios and corrections) and (vii) PubMed reference. Studies associating transplantation outcome and KIR were excluded, as they have heterogeneous designs - some studies associate KIR-ligand matches/mismatches based on recipients and donors samples, and others correlate the risk of relapse with KIR combinations. These study designs are under evaluation to ascertain if they can either fit the existing KDDB schema, or will be stored in a different database schema in the future.

The KIR and Diseases Database is part of Allele Frequencies Net Database, and can be accessed through the Allele Frequencies Net homepage (http://www.allelefrequencies.net/) using the menu 'KIR' and the submenu 'KIR and disease associations' or via direct URL access at http://www.allelefrequencies.net/diseases/. The interface to search and submit data is very similar to others previously implemented in AFND. The user can retrieve data from KDDB by the 'KIR and Diseases Database' link on KDDB homepage. All records can be filtered by gene (KIR gene or HLA ligand), country, continent or disease. Those filters can be applied alone or used in combination. The resulting page displays the data types mentioned previously. Additionally, the column 'World Distribution' contains links to the world map distribution of the associated KIR gene, and the column PubMed link redirects the user to the abstract in PubMed website (Figure 2).

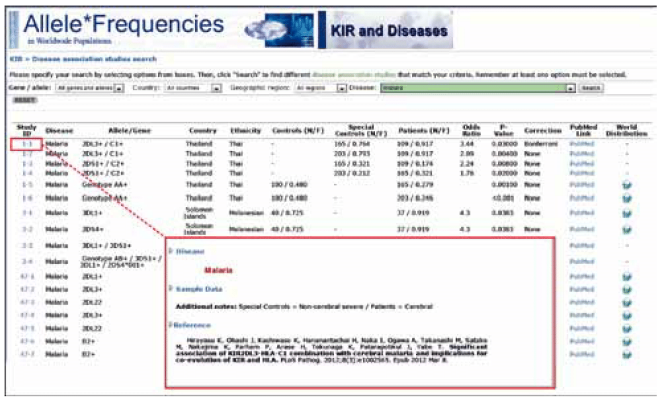


Figure 2: KDDB query page, showing the additional detail on a given association study retrieved by following the hyperlink.

The user can also submit studies to KDDB through the AFND homepage via the menu 'Submissions' and the submenu 'Add KIR and disease association study' or by accessing the 'KIR and Diseases Studies Submissions by Authors' link on the KDDB homepage. The data submission web form consists of four steps. The first step captures demographic information on the study including the number of patients and controls. Information is also captured on the geographic location of the population, the ethnicity and the bibliographic reference. The second step captures the disease association data – the genes, alleles, haplotypes, KIR-HLA ligands, etc., the disease name, the frequency of patients and controls exhibiting the given genetic profile and the results of the statistical test. The third step (optional) allows users to upload anonymised raw data (the KIR genetic profile of every individual in the study). The fourth step allows users to review their data and submit.

Although optional, capturing raw genotyping data for KIR (and also for HLA) would provide a powerful resource for immunogenetics research. For highly polymorphic loci, gene and allele frequencies are not enough to perform thorough analyses or ensure data quality. For example, having raw data available will allow analyses such as looking at disease associations in the centromeric or the telomeric regions. If the user has any problems uploading raw data through our system, he/she can alternatively send data direct to Derek.Middleton@rlbuht.nhs.uk for upload.

KDDB has the potential to be an important tool to aid the research community in understanding disease modulation exerted by the innate immune system, and can assist KIR researchers to review associations, perform meta-analysis or analyse combined data. Thus, it provides a new community resource for understanding not only how KIR genes are associated with disease, but also works in tandem with the large data sets already present in AFND, containing information on where particular disease associated KIR genes, genotypes or haplotypes are present in worldwide populations of different ethnic groups. Feedback is encouraged from the scientific community on the utility of the data submission and query interface and the general approach that has been taken to perform data curation.

This tool will only be as good as you, the scientific community, make it. Help is needed so that the website can be a useful addition to the Histocompatibility and Immunogenetics (H+I) field. Much of the frequency data sets in normal populations (submitted to AFND) have been taken from disease studies. We now encourage you to add data for the patients in the disease study. In addition to published studies, data from unpublished studies may also be added. Many studies are not published because they have not found an association, and by placing such data into the database, will correct this publication bias. It is anticipated that in the future having your results and study cited on a website will act as a reference that you can use in your CV. The H+I community is known throughout the world for collaboration. This is another occasion in which collaboration can help each other. If this venture is successful, the intention is to do something similar with HLA and disease associations. KIR-disease studies was initiated first as this field is relatively new, enabling most studies published or not, to be captured. Although it will be more difficult for HLA to obtain data from studies that were not published, the intention would be to provide a source for collecting such data on those diseases that have shown reproducible associations. This will be particularly important in diseases that have both KIR and HLA associations.

References

1. Ljunggren, H.G. and K. Karre, *In search of the 'missing self': MHC molecules and NK cell recognition.* Immunol Today, 1990. **11**(7): p. 237-44.
2. Middleton, D. and F. Gonzelez, *The extensive polymorphism of KIR genes.* Immunology, 2010. **129**(1): p. 8-19.
3. Gonzalez-Galarza, F.F., et al., *Allele frequency net: a database and online repository for immune gene frequencies in worldwide populations.* Nucleic Acids Res, 2011. **39**(Database issue): p. 9.
4. Takeshita, L.Y., et al., *A database for curating the associations between killer cell immunoglobulin-like receptors and diseases in worldwide populations.* Database, 2013. **12**(10).
5. Rogers, F.B., *Medical subject headings.* Bull Med Libr Assoc, 1963. **51**: p. 114-6.
6. Yu, W., et al., *A navigator for human genome epidemiology:* Nat Genet. 2008 Feb;40(2):124-5.

Transfusion Medicine and Hemotherapy

Review Article

Transfus Med Hemother 2014;41:352–355
DOI: 10.1159/000368056

Received: April 2, 2014
Accepted: September 3, 2014
Published online: September 16, 2014

Allele Frequencies Database

Louise Y.C. Takeshita[a] Andrew R. Jones[a] Faviel F. Gonzalez-Galarza[a] Derek Middleton[b]

[a]Institute of Integrative Biology, University of Liverpool, Liverpool, UK,
[b]Royal Liverpool University Hospital, Liverpool, UK

Keywords
Data management system · HLA · Immunogenetics · KIR

Summary
This review describes a database for the collection, archiving, sorting, searching and display of gene and allele frequencies for immunogenetic genes.

Introduction

The website Allele Frequencies Net Database (AFND) (*www.allelefrequencies.net*) was conceived and implemented in 2003, initially to collect and show frequencies of HLA alleles/allelic lineages, (thereafter in this publication called alleles), expressed as decimals, and the frequencies of individuals who carried the alleles, expressed as a percentage, in worldwide populations. Since then, data from other polymorphic immunogenetic loci, namely killer-cell immunoglobulin-like receptors (KIR) as well as MHC class I chain-related (MIC) and several cytokine polymorphisms, have been added [1]. Thus the database is highly relevant for those determining the alleles of all these immunogenetic genes, in solid organ and stem cell transplantation and in anthropology.

Demographics

To date AFND holds 1,324 populations differentiated by polymorphic region (table 1). In total data are available for 4,539,670 individuals. AFND provides an online repository with a set of querying tools to provide searching mechanisms. Data is collected in two aspects; demographic data to give details of the population and frequency data. The demographic data includes the number of individuals tested, the testing method used, the source of individuals e.g. anthropology study, the ethnicity, whether grandparents/parents came from same region and publication details, although data does not have to be published. Populations are given a name that best describes them; usually this consists of the name of country, region of country and if specified, ethnicity. If a new population cannot be differentiated by name from an existing population a number will be added, e.g. 'Japan 2'. If needed, 'Notes' are added for clarification. For example, if ambiguous data is added where a contributor has not been able to differentiate some alleles, a note is made to the fact that frequency has been added under the first allele of the ambiguity.

Submission of Data

The original intent of obtaining data was that individuals would directly, via an online submission system, add the demographic data via drop down boxes to the website. Thereafter, a file would be sent to them showing the HLA alleles to which the frequency data would be added (alleles from new releases of IMGT/HLA (*www.ebi.ac.uk/ipd/imgt/hla/*) are automatically addead to AFND, as are alleles that have their name changed). Unfortunately this mechanism of data entry has had limited success. Thus, more than 80% of the data has been added by us manually from peer-reviewed publications. As such AFND must be missing a lot of data which, although of good quality, is not published.

At present, we are working with the editor of *Human Immunology* to set up a procedure whereby frequency data can be published in a set format along with demographic data in a brief communication section of the journal. We believe this will aid our attempts to obtain more data directly from community submissions. In the future, with online data sets in-

KARGER

Fax +49 761 4 52 07 14
Information@Karger.com
www.karger.com

1660–3796/14/0415–0352$39.50/0

Accessible online at:
www.karger.com/tmh

Prof. Dr. Derek Middleton
Royal Liverpool University Hospital
Prescott Street, Liverpool, L7 8XP, UK
derek.middleton@rlbuht.nhs.uk

Table 1. Populations on AFND

Polymorphic region	Population studies	Gene/allele data	Haplotype data	Genotype data
HLA	925	910	365	–
KIR	225	224	–	143
Cytokine	114	114	–	–
MIC	60	60	21	–
Total	1,324	1,308	386	143

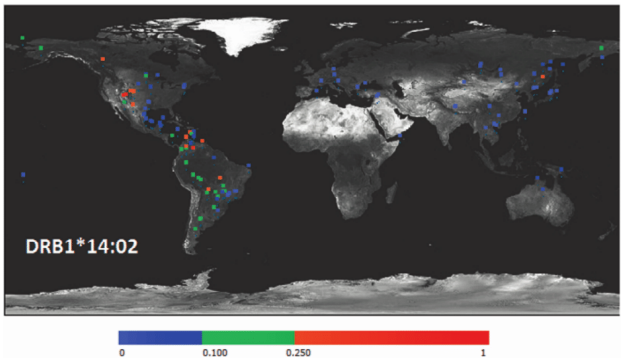


Fig. 1. Example of an allele frequency in worldwide populations.

creasingly replacing publications as a unit of citation (citation to a data set on a website acting as a peer-reviewed reference), an increase in the number of populations should ensue. In the intended format in *Human Immunology*, individuals will have to submit HLA types of the individuals making up their study. This will enable the data to be subjected to quality control aspects such as Hardy-Weinberg equilibrium testing and enabling meta-analysis to be run. It will also mean that existing data sets can be compared with other data sets at different time periods with regards to degree of resolution, new alleles being found, etc. It is envisaged that data sets submitted via this mechanism will have to fulfil minimum reporting requirements describing the method used for typing, the alleles tested for, how the allele interpretation was arrived at, etc.

AFND was set up so that alleles that could be determined but were not found were given a frequency of zero, to distinguish them from alleles that could not be determined in which case the frequency was left blank. If properly implemented, this would have meant that the website would show for a population which alleles had not been tested for, including those that would only be found and sequenced in later years. AFND was able to do this for data submitted directly but unfortunately most peer-reviewed publications do not describe which alleles could be determined and do not even give the release number from the IMGT/HLA database of the version of alleles available at the time of testing.

Searching the Data

Multiple filter schemas performed in each of the frequency searches allow users to optimise their searches and obtain matching results that best fit their requirements. Normally, the data can be searched according to the following filters: population, country, geographical region (e.g. West Europe) as well as level of resolution and source of data. Data can also be filtered on the sample size of population, by level of resolution and year of test. Apart from looking at all data for a specific locus, the user can search for specific alleles at that locus. Validation of data ensures that correct and current nomenclature is used. High resolution data is shown when low resolution alleles have been selected. For example, a search for HLA-A*01: 01 will also show allele frequencies, if available, at higher resolution that start with HLA-A*01: 01. The data can be displayed one population at a time, or one allele at a time in all populations and each of these categories can be sorted by highest to lowest frequency. Data can also be displayed on worldwide maps (fig. 1). The results of searches can be accessed programmatically via a URL. For example, the results of a search for global population frequencies for HLA-B*57: 01 can be accessed via the following address: *www.allelefrequencies.net/hla6006a.asp?hla_locus_type = Classical&hla_locus = B&hla_allele1 = B*57%3A01&hla_allele2 = B*57%3A01*. This mechanism thus enables other databases to

link to specific search results provided by AFND. AFND has also several links to other websites; for example, if an individual has performed a search and is interested in specific alleles, he/she can be directed specifically to details of that allele, including frequency, on IMGT/HLA.

HLA haplotype information is also provided. To date frequencies from 44,210 HLA haplotypes and 272 MICA-HLA-B associations from 3,838,492 individuals are available. When users wish to search for haplotype frequencies they may enter an allele at one or more loci, leaving other loci as 'any' or 'not' according to their needs. They will be provided with haplotype frequencies according to their search.

One of suggestions made to AFND has been to have more explicit reasoning behind how the data is displayed as some data could be open to misinterpretation. For example, in populations with 4-digit typing, AFND would, in addition to showing the 4-digit data, show 2-digit allele lineage data. Thus, it would appear that frequency of alleles added to more than one. In addition some data would not add up to 1, as not all allele frequencies were necessarily published. Thus, we are in process of adding an 'Explanation of Data' section whereby these points are elucidated and the populations they refer to are listed. Another example is that sometimes the numbers typed in the same population study for different loci are different. Previously, this was explained in the demographics but we also have thought it was worthwhile to show this in the 'Explanation of Data' section.

Archiving HLA Types

Initially HLA individual types making up a population study were not collected. However, such data is now being collected as part of an FP7 European grant, EUROSTAM. This grant seeks to assist individuals who, because of very high levels of sensitisation to HLA antigens, find it virtually impossible to receive a kidney transplant. The HLA laboratory at Eurotransplant led by F. Claas have shown that individuals with these high levels of antibodies (reacting with greater than 90% of the population) can receive a successful transplant by taking into account acceptable mismatches, which of course includes their own type. Despite this innovative idea, many patients still languish on their own country's transplant waiting list. The purpose of EUROSTAM is to ascertain in which other European country these individuals might have success in obtaining a transplant. Thus, AFND is now revisiting those populations that are already on the website along with new populations to obtain raw data, i.e. HLA types of individuals in these populations. Thus, it can be determined if some individuals would benefit from being on the transplant list of another European country.

Rare Alleles

The increase in number of HLA alleles, especially since molecular techniques came to the fore, is vast and continues to rise, even more so after Next Generation Sequencing is widely applied. This increase makes it very difficult for laboratories and vendors to keep their techniques current to determine these new alleles. At the last release from IMGT/HLA (release 3.15.0, January 2014) there are 10,533 known HLA alleles. A section in AFND has been introduced in an attempt to quantify the incidence of these alleles. Each allele is shown with information on whether the initial sequence submitted to IMGT/HLA has been confirmed and, if so, in how many individuals and in addition whether this allele has been found in individuals typed for the National Marrow Donor Programme (NMDP) or in individual laboratories. Most of the data from individual laboratories comes from projects conducted under auspices of the 15th and 16th International Histocompatibility and Immunogenetic Workshops [2, 3]. An individual searcher on AFND can determine their own criteria for what they want to use to define rarity of an allele. Previous reports during the International Workshops have shown that around 40% of HLA alleles have never been reported in another individual, after the report of the initially sequenced sample. Thus, laboratories can use this information in estimating what allele is present when they are faced with an ambiguous combination in the HLA type. At present AFND provides information on the country and ethnicity in which the allele is found, but the search is not able to show rare alleles filtered by each country or even in each geographic region, or filtered by ethnicity. This is one area where improvements are needed. But this depends totally on obtaining data from laboratories.

KIR Genes

Although the HLA section is probably the most used aspect of AFND, as mentioned previously data is also collected from other polymorphic immunogenetic gene systems. Of these KIR is the next most popular section of the database. Some additional data is supplied for KIR compared to HLA. Notably, frequencies of genotypes of KIR individuals (presence/absence of *KIR* gene) in different populations are available for 225 populations. To date 492 different genotypes are present on AFND, varying in their frequency; some being present in all populations studied, whereas others have only been found in one population, or indeed, in isolated cases, in one individual. When the genotype is found in only one population, it is conceivable that the data determined may not be accurate. Thus AFND provides a mechanism whereby the closest genotype to the rare genotype is displayed. This can indicate to the provider of the data, which gene could have been tested false-positive or false-negative in the unique gen-

otype. Each of the genotypes has been given a number, and it is gratifying for continuity purposes that in many publications authors list their genotypes according to that number.

The section for KIR has been augmented with a recently developed sub-section that shows KIR associations with disease [4]. It was thought that this was logistically practical as KIR determination in disease is a relatively new application, compared to HLA associations in disease, which have been reported for 50 years and would be too difficult to capture. One of the main purposes of this KIR/disease section is to collect non-published studies, which due to publication bias, are in the main studies that show no association.

Structural and Sequence-Based Analysis Tools

Another addition to AFND is the Amino Acid Section. Frequency data inputted at 4-digit level is converted to the frequency of individual amino acids at polymorphic positions in HLA alleles. In a disease association study, allele(s) can be shown to be more or less frequent in the disease cohort than the control population. However, in a way this is somewhat artificial, as it is not the allele as whole that is associated, but rather individual functional aspects of the allele. Thus, to examine the frequency of particular amino acids in disease and control cohorts is more meaningful. This section is a forerunner to the new section on HLA epitopes, which was stimulated by the collection of data from the EUROSTAM project. In transplantation, HLA epitope data is starting to change the current view of HLA matching, from allele matching to structural matching where epitopes are patches of polymorphic residues that can stimulate production of specific anti-HLA antibodies, a concept especially important in preventing high sensitization of transplant patients [5]. The new HLA epitopes part of the site uses the nomenclature released through the HLA Epitope Registry [6], which indicates the mapping from HLA allele-level nomenclature to confirmed or predicted epitopes. In AFND the epitope section shows the frequencies of these epitopes in different populations, with various ways for querying and visualisation. Initial data in worldwide populations is being delivered from 4-digit HLA data, either from raw data or by statistical means from haplotype data.

Conclusion

AFND now contains the most extensive archive of immune gene/allele frequencies. The success can be judged by the 300 hits received on average every day. AFND has been used in many applications. Although not comparable to actually searching on stem cell registries, it can be used to give some indication of which country a potential donor might be found, particularly if one uses the haplotype information available. It has also been used extensively in clinical laboratories to aid in determining HLA types and in anthropology research units for comparative purposes. However, for AFND to be a continued success we rely on help from colleagues. AFND is very willing to listen to feedback, suggestions and criticisms of the website, in order for improvement to ensue. One area in which the readers of this review might be interested is in whether there should be a section on blood group frequencies in worldwide populations. To do this, we would need to liaise and obtain assistance from someone familiar to this field. This might be suitable for e.g. a Master's project for anyone interested.

In the future it is our intention to have available statistical packages on AFND so that these can be used by individuals to examine their data, to compare them with other populations or to perform analysis on existing populations of their choice.

Disclosure Statement

None of the authors have a conflict of interest.

References

1 Gonzalez-Galarza FF, Christmas S, Middleton D, Jones AR: Allele frequency net: a database and online repository for immune gene frequencies in worldwide populations. Nucleic Acids Res 2011; 39:D913–919.

2 Gonzalez-Galarza FF, Mack SJ, Hollenbach J, Fernandez-Vina M, Setterholm M, Kempenich J, Marsh SG, Jones AR, Middleton D; HLA Rare Allele Consortium: 16th IHIW: extending the number of resources and bioinformatics analysis for the investigation of HLA rare alleles. Int J Immunogenet 2013; 40: 60–65.

3 Middleton D, Gonzalez F, Fernandez-Vina M, Tiercy JM, Marsh SG, Aubrey M, Bicalho MG, Canossi A, Carter V, Cate S, Guerini FR, Loiseau P, Martinetti M, Moraes ME, Morales V, Perasaari J, Setterholm M, Sprague M, Tavoularis S, Torres M, Vidal S, Witt C, Wohlwend G, Yang KL: A bioinformatics approach to ascertaining the rarity of HLA alleles. Tissue Antigens 2009; 74: 480–485.

4 Takeshita LY, Gonzalez-Galarza FF, dos Santos EJ, Maia MH, Rahman MM, Zain SM, Middleton D, Jones AR: A database for curating the associations between killer cell immunoglobulin-like receptors and diseases in worldwide populations. Database (Oxford) 2013; 2013:bat021.

5 Duquesnoy RJ: Antibody-reactive epitope determination with HLAMatchmaker and its clinical applications. Tissue Antigens 2011; 77: 525–534.

6 Duquesnoy RJ, Marrari M, da M Sousa LC, de M Barroso JR, de S U Aita KM, da Silva AS, do Monte SJ: 16th IHIW: a website for antibody-defined HLA epitope Registry. Int J Immunogenet 2013; 40: 54–59.

D784–D788 *Nucleic Acids Research, 2015, Vol. 43, Database issue* *Published online 20 November 2014*
doi: 10.1093/nar/gku1166

Allele frequency net 2015 update: new features for HLA epitopes, KIR and disease and HLA adverse drug reaction associations

Faviel F. González-Galarza[1,2,†], Louise Y.C. Takeshita[1,†], Eduardo J.M. Santos[1,3], Felicity Kempson[1], Maria Helena Thomaz Maia[3], Andrea Luciana Soares da Silva[3], André Luiz Teles e Silva[3], Gurpreet S. Ghattaoraya[1,4], Ana Alfirevic[4], Andrew R. Jones[1,*] and Derek Middleton[5,6]

[1]Institute of Integrative Biology, University of Liverpool, Liverpool, UK, [2]Center for Biomedical Research, Faculty of Medicine, Autonomous University of Coahuila, Torreon, Mexico, [3]Human and Medical Genetics, Institute of Biological Sciences, Federal University of Pará, Brazil, [4]Department of Molecular and Clinical Pharmacology, Institute of Translational Medicine, University of Liverpool, Liverpool, UK, [5]Transplant Immunology Laboratory, Royal Liverpool and Broadgreen University Hospital, University of Liverpool, UK and [6]Institute of Infection and Global Health, University of Liverpool, UK

Received September 23, 2014; Revised October 27, 2014; Accepted October 30, 2014

ABSTRACT

It has been 12 years since the Allele Frequency Net Database (AFND; http://www.allelefrequencies.net) was first launched, providing the scientific community with an online repository for the storage of immune gene frequencies in different populations across the world. There have been a significant number of improvements from the first version, making AFND a primary resource for many clinical and scientific areas including histocompatibility, immunogenetics, pharmacogenetics and anthropology studies, among many others. The most widely used part of AFND stores population frequency data (alleles, genes or haplotypes) related to human leukocyte antigens (HLA), killer-cell immunoglobulin-like receptors (KIR), major histocompatibility complex class I chain-related genes (MIC) and a number of cytokine gene polymorphisms. AFND now contains >1400 populations from more than 10 million healthy individuals. Here, we report how the main features of AFND have been updated to include a new section on 'HLA epitope' frequencies in populations, a new section capturing the results of studies identifying HLA associations with adverse drug reactions (ADRs) and one for the examination of infectious and autoimmune diseases associated with KIR polymorphisms—thus extending AFND to serve a new user base in these growing areas of research. New criteria on data quality have also been included.

INTRODUCTION

The Allele Frequency Net Database (AFND) was designed to provide a free centralized resource for the storage of frequencies on the polymorphisms of several immune-related genes (1). The website contains information primarily on the frequencies of several genes from the human leukocyte antigens (HLA) system, killer-cell immunoglobulin-like receptors (KIR), major histocompatibility complex class I chain-related genes (MIC) and a number of cytokine gene polymorphisms. These loci are among the most polymorphic in humans and play key roles in the immune system response, as well as being important for donor-recipient matching in organ and stem cell transplantation success (2,3). These loci have also been studied extensively due to associations between polymorphisms and response to infectious diseases (4) or susceptibility to autoimmune diseases (5–7). Recently, there is also a growing field of study identifying associations between particular HLA polymorphisms and increased risk for adverse drug reactions (ADRs) (8,9). The HLA region is also commonly analyzed in anthropology studies (10). The HLA system comprises more than 20 genes, however, only six loci are routinely typed by laboratories, i.e. HLA-A, -B, -C for Class I and HLA-DRB1, -DQB1 and -DPB1 for Class II. Hence, most of the data sets in AFND cover principally these genes, also known as classical HLA loci. At present, more than 11 000 HLA class I

*To whom correspondence should be addressed. Tel: +44 151 795 4514; Email: Andrew.Jones@liv.ac.uk
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

or II alleles have been reported at the IMGT/HLA database (Release 3.17.0.1, August 2014) (11).

The first release of AFND in 2003 consisted of only a few sections, and frequencies of HLA alleles/allelic lineages were shown in static web pages. In 2008, the database was substantially re-developed, producing the system described in a previous publication in the 2011 database issue of *Nucleic Acid Research* (1), which readers should consult for a detailed description of the purpose and background to the system. Since then, the database has grown substantially in terms of the number of populations covered and the number of users/citations. In the past 3 years, more than 75 000 different users from 172 countries accessed the database. In this article, we describe new population data added, new developments in validating the quality of data sets in AFND, as well as new sections for capturing frequency data on 'HLA epitopes' (structure-level polymorphisms recognized by antibodies), associations between KIR polymorphisms and disease and associations between HLA alleles and ADRs that have been identified from the literature.

DESCRIPTION OF AFND AND SOURCES OF DATA

The database of AFND is currently implemented in MS SQL Server in the latest release (previously maintained in MySQL). Web pages are constructed using active server pages, Javascript and AJAX technology to improve user interaction and data visualization.

Normal population data

AFND receives data from three main sources: (i) data from peer-reviewed publications, (ii) from populations that are analyzed at International HLA and Immunogenetics Workshops (IHWS) and (iii) submissions from individual laboratories across the world. However, by far the most data (80%) come from data extraction and curation by the AFND team from peer-reviewed publications. As such, a vast amount of data may be missing, which, although of good quality, is not published and we encourage labs with such data to contact us. The literature review comprises not only histocompatibility- and immunogenetics-related journals, but also, we have established semi-automated methods using regular structured queries of literature databases to verify other journals that may contain suitable data for inclusion.

As of September 2014, we have collected information on >1400 healthy populations from more than 10 million people. The HLA section contains the majority of the submissions with 1022 populations, followed by populations analyzed for polymorphisms in KIR (229), cytokine genes (114) and MIC (60) (Table 1, figures correct in September 2014). Currently, data sets from 138 countries are included within AFND—with highest coverage (by population number) the United States (121 populations), followed by China (110 populations)—summarized under the 'Populations-Pops By Region' menu in the database. In terms of the number of individuals, United States, Brazil and Italy have the largest amount of data, due to the inclusion of large data sets from bone marrow donor registries in the database. As described previously (1), the most popular tools in AFND include queries for particular allele/haplotype frequencies (viewed as a table or world map—Figure 1A), or analysis of all allele/haplotype frequencies within a given population or geographic region of the world.

Although all submissions by contributors are considered for inclusion, AFND has introduced minimal criteria before the population becomes publicly accessible on the website. These minimum requirements include the homogenization on the naming of the populations, an appropriate assignment of the geographical region to which the population belongs, validation of frequency data such as ensuring allele names comply with the official nomenclature guidelines as described at http://allelefrequencies.net/quality.asp. These guidelines will continue to develop and be implemented across all data sets newly added to AFND.

HLA epitope database

The presence of anti-HLA donor-specific antibodies in transplant patients is a crucial factor related to tissue and graft rejection. These antibodies target specific regions of HLA proteins that are different from the transplant patient's HLA proteins—termed 'HLA epitopes'. Current efforts in matching for kidney transplantation minimizes the number of HLA antigen mismatches (very rarely a perfect match is achieved), yet this matching disregards structural differences (or similarities) between HLA proteins. As this concept is starting to get more recognition, and epitopes are being systematically defined (12), we developed a new section within AFND called Epitope Frequency Database (EpFreq-DB), for the storage of HLA epitope frequencies (the percentage of the population expressing a given epitope) across worldwide populations.

Two sources of data were used to generate HLA epitope frequencies: (i) HLA haplotype frequency data from AFND and (ii) HLA raw genotyping data, in both cases using data sets with at least 4-digit resolution (e.g. A*01:01). In this context 'raw data' is the genotype, comprising one or two alleles called at each HLA locus, of every individual in the population. Low-resolution data (2 digits) can encompass alleles with differences in the protein sequence, and thus epitopes cannot be unambiguously determined. Allele frequency data sets cannot be used for accurate inference of HLA epitope frequencies because the same epitope can be present in alleles at different HLA loci of the same individual, e.g. epitopes shared between some combination of HLA-A, B and C genes.

For calculating epitopes frequencies, two different methods were used according to the data type. From HLA raw genotyping data, they were calculated by counting individuals having at least one allele expressing a given epitope. The number of individuals expressing the epitope is then divided by the population sample size. From HLA haplotype frequency data, the Hardy–Weinberg equilibrium calculation ($p^2 + 2pq + q^2$) was applied to estimate epitope frequencies, treating a haplotype as expressing a given allele p or not q. The method has been extensively validated, and produces highly accurate estimates of epitope frequencies (e.g. $r^2 \sim = 0.99$ versus estimates from raw data). A full description of the methodology will follow in a subsequent manuscript.

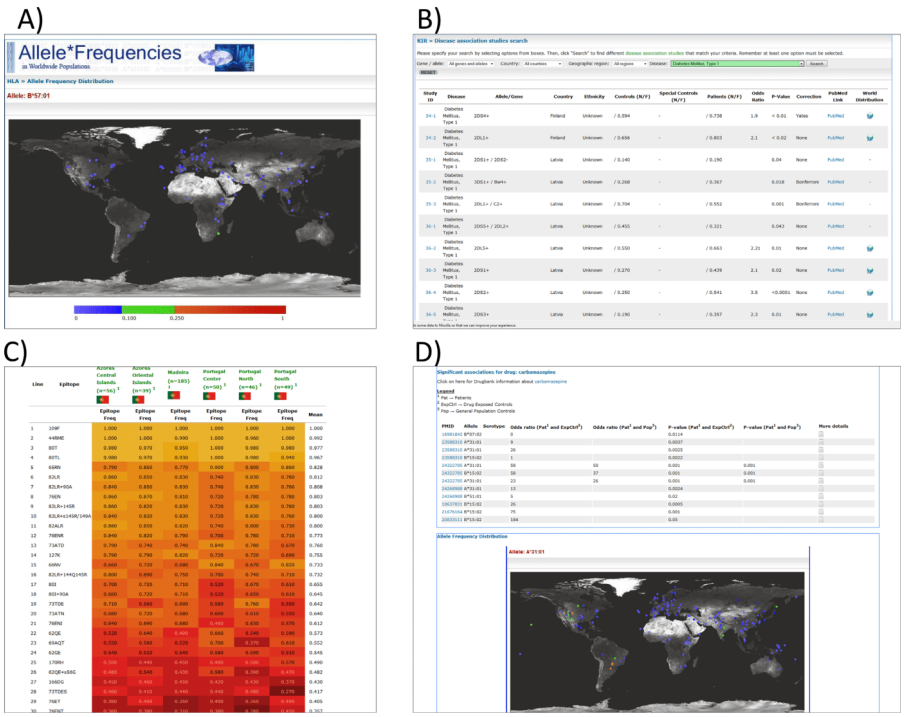


Figure 1. (**A**) A world map in AFND showing the global allele frequencies of HLA B*57:01, which, for example, has been associated with adverse reactions to abacavir; (**B**) a query of KDDB for populations/studies in which associations between KIR genotypes have been made with Type 1 diabetes; (**C**) a heat map view of several populations in the HLA epitope database (filtered by epitopes from 'locus A+B'); (**D**) a drug report in AFND showing all association data for carbamazepine.

Table 1. Frequency data sets by polymorphic region at AFND, figures correct in September 2014

Polymorphic region	Population studies	Gene/allele data	Haplotype data	Genotype data
HLA	1022	1004	370	-
KIR	229	228	-	146
Cytokine	114	114	-	-
MIC	60	60	21	-
Totals	1425	1406	391	146

To date, the HLA Epitope Frequencies Database (EpFreq-DB) comprises HLA class I epitope frequencies from 41 worldwide populations (17 raw/genotype data sets and 24 haplotype data sets) comprising more than 36 000 individuals. The definition/nomenclature of HLA epitopes and corresponding HLA alleles used in EpFreq-DB is based on the HLA Epitope Registry (http://www.epregistry.ufpi.br/) (12). EpFreq-DB has a query page for analyzing the epitope frequencies according to a number of different criteria or filters, displaying results as a tabular view, projected onto world maps for single epitopes, or as a comparison of epitope frequencies in different populations as a heatmap (Figure 1C).

KIR and diseases

The section for KIR now includes a recently developed section (KIR and Disease Database—KDDB) containing associations that have been identified in the literature between KIR polymorphisms and disease—for detailed discussion and methods see (13). The development of KDDB was initi-

ated, since there is a growing area of research demonstrating that the KIR genes carried by an individual can increase or decrease risk/severity of auto-immune and infectious diseases. However, many studies have relatively small sample sizes and different studies have conflicting findings. The development of KDDB enables researchers to examine all the published studies in one place, for example to foster meta-analyses and determine if findings in one study have been confirmed elsewhere.

Currently, KDDB has a total of 1179 KIR disease-association determinants captured from 204 articles, including those with single KIR genes, profiles of combined KIR genes and/or HLA class I ligands, and full KIR genotypes. According to the present database, KIR associations of 79 different disease terms have been included, of which 19 associations can be classified as infectious diseases, 32 as autoimmune or idiopathic diseases, three related to pregnancy, 16 to cancer, eight to chronic inflammatory diseases and one mental disorder. The web interface allows users to query KIR and disease associations applying several filters related to population demographics, disease studies and gene features. From the same location, the user can access links to submit new studies to the database, for example including those studies in which no association has been found—which are difficult to publish (and thus otherwise contribute to publication bias).

ADR database

One of the biggest problems faced by clinicians and the pharmaceutical companies is the risk that patients might experience ADRs upon exposure to a drug treatment. Approximately 10% of all ADRs are immune mediated (14) and the most significant genetic associations have been related to HLA alleles (8). Given the huge inter-individual variability in HLA alleles only a small number of individuals are reported in each study leading to statistical analysis with low power. To assist the HLA and pharmacogenetic community, we have collated data sets from the literature, and they can be queried alongside the large data collection for normal populations within AFND at the allele and haplotype level. This provides a resource that not only facilitates meta-analyses but also enables users to examine the quality of published studies by comparing the frequencies of HLA alleles in 'control' cohorts with worldwide populations.

A similar curation protocol to KDDB was followed. Two inclusion criteria were used: first, the included studies utilized a case-control design, which provided statistical evidence for the association; second, high-resolution HLA typing was performed to generate data (low-resolution data are only present in the database for studies that performed both low- and high-resolution typing). Low-resolution data sets may be included in a later release of the database if we see demand from users for their inclusion. We included information on ethnicity, drug of interest and proportion of cases and controls that carry the HLA allele implicated in ADRs. Associations with >20 different drugs are captured in the current beta release of the database, with anti-epileptic drug carbamazepine having the most studies included. The aim of this new feature of AFND is under active development, and we aim to cover all published studies, and, as such, the amount of data included and the query tools provided will increase over the coming years. We have developed a feature—called 'Drug reports' highlighting all known associations for a given drug (Figure 1D). Users can see the worldwide distributions for the implicated alleles and haplotype data (from healthy data sets in AFND) and links out to IMGT/HLA for sequence alignments of the implicated alleles.

FUTURE DEVELOPMENTS

Future challenges and plans for AFND include improving the direct data submission process for all sections within AFND, for example by a direct connection with the Human Immunology journal as part of the manuscript submission process. We will also continue to develop quality control tools, for example developing a 'gold standard' set of populations in AFND that meet high quality/validation criteria. We also plan to develop the ability to perform statistical analyses and visualization of population data on the AFND site to facilitate users and as an additional mechanism for stimulating direct data submissions. To do this we will be asking that submissions include raw data.

CONCLUSION

AFND is the most comprehensive database for frequency data, relating to immune genes/alleles in worldwide populations. AFND receives ∼300 hits per day, and is widely cited in a variety of different clinical and research fields. The origins of the database were to provide support for the Histocompatibility and Immunogenetics community, in understanding worldwide distribution of HLA genotypes. The database has continued to support these fields, while developing new features (such as HLA epitope frequencies) that provide a new viewpoint on these highly complex data sets. AFND has recently expanded to support new research groups, particularly those working on autoimmune disorders and infectious diseases, for which associations with KIR polymorphism are increasingly being identified. AFND is also developing new features to support pharmacogenetic research, since there is a rapidly growing field emerging, as it becomes increasingly clear that HLA molecules play a role in many ADRs.

AVAILABILITY

AFND Homepage: http://www.allelefrequencies.net
Contact: support@allelefrequencies.net

ACKNOWLEDGEMENTS

We would like to thank Steven J. Mack, Jill Hollenbach, Jose Nunes and Alicia Sanchez-Mazas for their inputs to the data quality section; Rene Duquesnoy, Luiz Cláudio Demes da Mata Sousa and Semiramis Jamil Hadad do Monte and colleagues for assistance on the HLA epitopes section; and to all contributors who have provided data to this database.

FUNDING

Science without borders program - CNPq (National Council for Scientific and Technological Development—Brazil) [PhD studentship to L.Y.C.T. and Post-Doctoral Grant grant to E.J.M.S.]; National Council on Science and Technology of Mexico [FOINS-CONACYT-217830-R to F.F.G.G.]. Funding for open access charge: University of Liverpool Library.
Conflict of interest statement. None declared.

REFERENCES

1. Gonzalez-Galarza,F.F., Christmas,S., Middleton,D. and Jones,A.R. (2011) Allele frequency net: a database and online repository for immune gene frequencies in worldwide populations. *Nucleic Acids Res.*, **39**, D913–D919.
2. Opelz,G. and Dohler,B. (2013) HLA matching and kidney transplantation: beyond graft survival. In: Everly,MJ and Terasaki,PI (eds). *Clinical Transplants 2013*. Terasaki Foundation Laboratory, Los Angeles, CA, pp. 121–126.
3. Nowak,J. (2008) Role of HLA in hematopoietic SCT. *Bone Marrow Transplant.*, **42**(Suppl. 2), S71–S76.
4. Blackwell,J.M., Jamieson,S.E. and Burgner,D. (2009) HLA and infectious diseases. *Clin. Microbiol. Rev.*, **22**, 370–385.
5. Bluestone,J.A., Herold,K. and Eisenbarth,G. (2010) Genetics, pathogenesis and clinical interventions in type 1 diabetes. *Nature*, **464**, 1293–1300.
6. Klein,J. and Sato,A. (2000) The HLA system. Second of two parts. *N. Engl. J. Med.*, **343**, 782–786.
7. Thorsby,E. and Lie,B.A. (2005) HLA associated genetic predisposition to autoimmune diseases: genes involved and possible mechanisms. *Transpl. Immunol.*, **14**, 175–182.
8. Alfirevic,A. and Pirmohamed,M. (2010) Drug induced hypersensitivity and the HLA complex. *Pharmaceuticals*, **4**, 69–90.
9. Yip,V.L., Marson,A.G., Jorgensen,A.L., Pirmohamed,M. and Alfirevic,A. (2012) HLA genotype and carbamazepine-induced cutaneous adverse drug reactions: a systematic review. *Clin. Pharmacol. Ther.*, **92**, 757–765.
10. Abi-Rached,L., Jobin,M.J., Kulkarni,S., McWhinnie,A., Dalva,K., Gragert,L., Babrzadeh,F., Gharizadeh,B., Luo,M., Plummer,F.A. *et al.* (2011) The shaping of modern human immune systems by multiregional admixture with archaic humans. *Science*, **334**, 89–94.
11. Robinson,J., Halliwell,J.A., McWilliam,H., Lopez,R., Parham,P. and Marsh,S.G. (2013) The IMGT/HLA database. *Nucleic Acids Res.*, **41**, D1222–D1227.
12. Duquesnoy,R.J., Marrari,M., da M. Sousa,L.C., de M. Barroso,J.R., de S.U. Aita,K.M., da Silva,A.S. and do Monte,S.J. (2013) 16th IHIW: a website for antibody-defined HLA epitope Registry. *Int. J. Immunogenet.*, **40**, 54–59.
13. Takeshita,L.Y.C., Gonzalez-Galarza,F.F., dos Santos,E.J.M., Maia,M.H.T., Rahman,M.M., Zain,S.M.S., Middleton,D. and Jones,A.R. (2013) A database for curating the associations between killer cell immunoglobulin-like receptors and diseases in worldwide populations. *Database*, **2013**, bat021.
14. Chen,C.J., Cheng,C.F., Lin,H.Y., Hung,S.P., Chen,W.C. and Lin,M.S. (2012) A comprehensive 4-year survey of adverse drug reactions using a network-based hospital system. *J. Clin. Pharm. Ther.*, **37**, 647–651.

J Antimicrob Chemother 2017; **72**: 1152–1162
doi:10.1093/jac/dkw545 Advance Access publication 5 January 2017

Journal of Antimicrobial Chemotherapy

Genome-wide association study of nevirapine hypersensitivity in a sub-Saharan African HIV-infected population

Daniel F. Carr[1]*†, Stephane Bourgeois[2]†, Mas Chaponda[1,3], Louise Y. Takeshita[4], Andrew P. Morris[1,5], Elena M. Cornejo Castro[1], Ana Alfirevic[1], Andrew R. Jones[4], Daniel J. Rigden[4], Sam Haldenby[6], Saye Khoo[1], David G. Lalloo[7], Robert S. Heyderman[3,8], Collet Dandara[9], Elizabeth Kampira[9], Joep J. van Oosterhout[3,10], Francis Ssali[11], Paula Munderi[12], Giuseppe Novelli[13], Paola Borgiani[13], Matthew R. Nelson[14], Arthur Holden[15], Panos Deloukas[2,16,17] and Munir Pirmohamed[1]

[1]Department of Molecular and Clinical Pharmacology, University of Liverpool, Liverpool, UK; [2]William Harvey Research Institute, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London, UK; [3]Malawi-Liverpool-Wellcome Trust Clinical Research Programme, College of Medicine, University of Malawi, Malawi; [4]Institute of Integrative Biology, University of Liverpool, Liverpool, UK; [5]Department of Biostatistics, University of Liverpool, Liverpool, UK; [6]Centre for Genomic Research, University of Liverpool, Liverpool, UK; [7]Liverpool School of Tropical Medicine, Liverpool, UK; [8]Division of Infection and Immunity, University College London, London, UK; [9]Division of Human Genetics, University of Cape Town, Cape Town, South Africa; [10]Dignitas International, Zomba, Malawi; [11]Joint Clinical Research Centre, Headquarters, Kampala, Uganda; [12]UVRI/MRC Uganda Research Unit on AIDS, Entebbe, Uganda; [13]Department of Biomedicine and Prevention, University of Rome 'Tor Vergata', Rome, Italy; [14]GlaxoSmithKline, Research Triangle Park, NC, USA; [15]SAEC Consortium, Ltd, Chicago, IL, USA; [16]Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK; [17]Princess Al-Jawhara
Al-Brahim Centre of Excellence in Research of Hereditary Disorders (PACER-HD), King Abdulaziz University, Jeddah, 21589, Saudi Arabia

*Corresponding author. Tel: +44-151-795-5392; E-mail: d.carr@liv.ac.uk
†Authors contributed equally.

Received 15 August 2016; returned 10 October 2016; revised 13 October 2016; accepted 20 November 2016

Background: The antiretroviral nevirapine is associated with hypersensitivity reactions in 6%–10% of patients, including hepatotoxicity, maculopapular exanthema, Stevens–Johnson syndrome (SJS) and toxic epidermal necrolysis (TEN).

Objectives: To undertake a genome-wide association study (GWAS) to identify genetic predisposing factors for the different clinical phenotypes associated with nevirapine hypersensitivity.

Methods: A GWAS was undertaken in a discovery cohort of 151 nevirapine-hypersensitive and 182 tolerant, HIV-infected Malawian adults. Replication of signals was determined in a cohort of 116 cases and 68 controls obtained from Malawi, Uganda and Mozambique. Interaction with *ERAP* genes was determined in patients positive for *HLA-C*04:01*. *In silico* docking studies were also performed for *HLA-C*04:01*.

Results: Fifteen SNPs demonstrated nominal significance ($P < 1 \times 10^{-5}$) with one or more of the hypersensitivity phenotypes. The most promising signal was seen in SJS/TEN, where rs5010528 (*HLA-C* locus) approached genome-wide significance ($P < 8.5 \times 10^{-8}$) and was below *HLA*-wide significance ($P < 2.5 \times 10^{-4}$) in the meta-analysis of discovery and replication cohorts [OR 4.84 (95% CI 2.71–8.61)]. rs5010528 is a strong proxy for *HLA-C*04:01* carriage: *in silico* docking showed that two residues (33 and 123) in the B pocket were the most likely nevirapine interactors. There was no interaction between *HLA-C*04:01* and *ERAP1*, but there is a potential protective effect with *ERAP2* [$P = 0.019$, OR 0.43 (95% CI 0.21–0.87)].

Conclusions: *HLA-C*04:01* predisposes to nevirapine-induced SJS/TEN in sub-Saharan Africans, but not to other hypersensitivity phenotypes. This is likely to be mediated via binding to the B pocket of the HLA-C peptide. Whether this risk is modulated by *ERAP2* variants requires further study.

Introduction

Nevirapine, an NNRTI used for HIV[1] infection is effective[2] as part of combination antiretroviral therapy, but causes hypersensitivity in 6%–10% of patients.[3,4] This can manifest in various ways, ranging from nevirapine-induced rash (NIR) (i.e. a maculopapular exanthema without any systemic manifestations), hypersensitivity syndrome (HSS) to severe blistering skin reactions such as Stevens–Johnson syndrome (SJS) and toxic epidermal necrolysis (TEN)[5] (1–2 per 1000 exposed individuals[6]). Extra-cutaneous involvement typically manifests as hepatotoxicity.[7]

Identification of the genetic risk factors for nevirapine hypersensitivity has focused on candidate gene approaches. Nevirapine is primarily metabolized by the hepatic cytochrome P450s 2B6 (CYP2B6) and 3A4 (CYP3A4).[8] The exon 4 variant in *CYP2B6* (c.516G>T), which encodes a non-synonymous amino acid substitution (Gln172His) (rs3745274), leads to loss of function,[9,10] with the variant T allele resulting in higher nevirapine plasma concentrations in both Caucasian[11] and sub-Saharan[12] adult patients. The associations with *CYP2B6* polymorphisms are rather confusing with the *CYP2B6* c.516G>T SNP associated with nevirapine-induced cutaneous adverse events in black and white populations[13] but not with nevirapine-induced hepatotoxicity.[14] The association with HLA alleles is even more complex, with *HLA-DRB1*01:01* (Caucasian[13,15,16]), *HLA-C*04* (Thai,[17] Chinese[18] and Black[13]), *HLA-C*08* (Japanese[19]) and *HLA-B*35:05* (Thai[13,20]) acting as predisposing alleles for nevirapine hypersensitivity. Our own previous study within a subset of patients from the Malawian HIV population described in this paper identified an association between *HLA-C*04:01* and nevirapine-induced SJS.[21]

In this study, in order to overcome some of the issues associated with candidate gene analysis, we have undertaken a genome-wide association study (GWAS) in a Malawian HIV cohort of nevirapine-exposed patients in order to identify genetic biomarkers of nevirapine hypersensitivity in an unbiased manner. We have also investigated whether there is any interaction between *HLA-C*04:01* in SJS/TEN patients and the endoplasmic reticulum aminopeptidase genes (*ERAP1* and *ERAP2*), which have been shown to modulate the risk of various immune diseases, in particular ankylosing spondylitis.[22]

Methods

Patients

Discovery cohort

Antiretroviral-naive patients ($n = 1117$) were prospectively recruited as previously described[21] (Figure 1) from the Queen Elizabeth Central Hospital (QECH), Blantyre, Malawi, between March 2007 and December 2008. All were self-reported black African, over the age of 16, and had no baseline jaundice. CD4+ counts and liver function tests were monitored at 0, 2, 6, 10, 14, 18 and 22 weeks. Fifty-seven patients from this prospective cohort had nevirapine-induced hypersensitivity fulfilling the criteria of one or more of the following phenotypes:

- NIR: widespread maculopapular exanthema with no systemic manifestations but which worsened on treatment continuation.
- HSS: widespread rash with systemic manifestations (i.e. fever, cough or abnormal liver function tests). This is also known as DRESS (drug reaction with eosinophilia and systemic symptoms).
- SJS: blistering eruption affecting <10% of body surface area with two or more mucous membranes involved.
- TEN: blistering rash affecting >30% of body surface area and two or more mucous membranes. Patients with overlap syndrome had 10%–30% of their body surface area affected.
- Drug-induced liver injury (DILI): jaundice and abnormal ALT.

In addition, a total of 149 cases of nevirapine-induced hypersensitivity were recruited prospectively from QECH separately from the study described above, and a further 28 were identified retrospectively from patient records at the same centre. Out of a total of 234 hypersensitive cases, 159 where sufficient genomic DNA was available were included, along with 193/1060 of the nevirapine-treated age- and gender-matched controls (352 in total), in the discovery GWAS. Numbers of tolerant controls included were constrained by DNA quality and quantity.

Replication cohort

We recruited a number of patients with nevirapine hypersensitivity, with different phenotypes, from a number of centres (Table 1) to replicate our findings:

- Thirty nevirapine-hypersensitive patients and matched (age and gender) HIV-positive nevirapine-treated controls from Malawi. All controls and eight of the cases were from the original study but not included in the initial GWAS due to DNA quantity restraints. The other 22 cases presenting with the hypersensitivity phenotype according to the above criteria were identified from the QECH after the conclusion of the initial recruitment phase (December 2008).
- Thirty-two nevirapine-hypersensitive cases and age- and gender-matched controls identified in Uganda from the DART study cohort.[23] Cases were defined according to available patient records and subsequently categorized into the sub-phenotypes described above.
- Twenty-seven pregnant female patients with nevirapine-induced hepatotoxicity and 10 nevirapine-tolerant pregnant controls from Mozambique. Cases were defined as previously stated,[24] and included patients who discontinued nevirapine due to increased liver enzymes (grade 3/4). Controls were excluded if ALT/AST levels exceeded the median value observed in the case cohort.
- Twenty-seven female patients with nevirapine-induced SJS/TEN from Mozambique.[25] In this instance SJS/TEN was defined as development of exanthema and blistering starting mainly on the trunk, involving ≥10% of the body surface with mucosal involvement.

Ethics

Full ethics approval for the study was received from the Liverpool School of Tropical Medicine Research Ethics Committee (Liverpool, UK), the College of Medicine Research and Ethics Committee, University of Malawi (Blantyre, Malawi) and the Uganda National Council for Science and Technology. All patients gave their written informed consent and those who met the criteria for a case had nevirapine withdrawn in accordance with Malawian National Treatment Guidelines. Local ethics approval was obtained for the DART study as previously described[23] with subsequent ethics approval for a pharmacogenetic sub-study also obtained.[26]

DNA extraction

Genomic DNA was extracted from whole blood for the discovery cohort[21] and replication cohorts as previously described.[24,25]

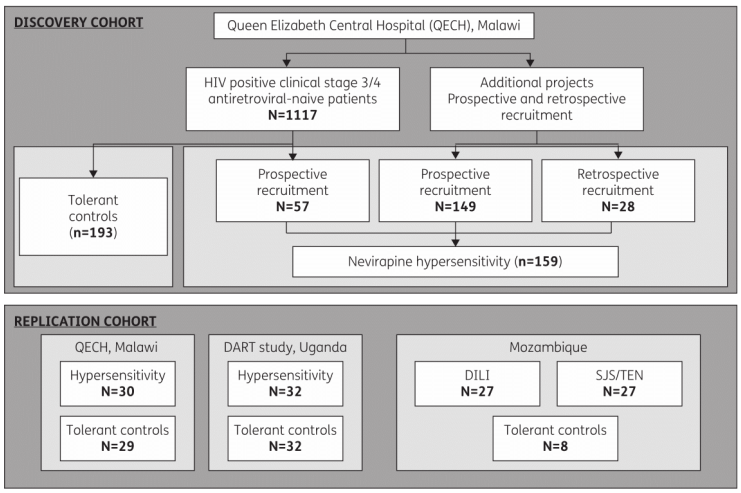


Figure 1. Schematic representation of the source of both nevirapine-hypersensitive and -tolerant patients for the GWAS discovery and replication cohorts.

Discovery cohort genotyping and sample quality control (QC)

A total of 352 samples were genotyped for 1 048 713 variants using the HumanOmni1-Quad_v1 chip (Illumina). Variants were excluded from analysis if their minor allele frequency (MAF) was <1%, the call rate was <99% for an MAF between 1% and 3%, the call rate was <98% for an MAF of >3%, or if Hardy–Weinberg expectations were not satisfied ($P < 10^{-4}$).

Individuals were excluded if the sample call rate was <95%, the assigned gender contradicted genetic information from the X chromosome heterozygosity, or if they appeared to be duplicates, or related to other individuals in the study (as measured by identity by state using PLINK[27]). Multidimensional scaling analysis of genotype data was undertaken by merging the data with HapMap3 cohort data and using the mds function in PLINK[27] in order to determine population stratification (Figure S1, available as Supplementary data at *JAC* Online).

Discovery cohort imputation

Imputation of genotypes, after phasing of each chromosome using ShapeIt,[28] was carried out using IMPUTE V2.3.1,[29] 1000G phase 1 integrated v3 macGT1 reference panel haplotypes (March 2012).[30] After imputation, SNPs with an information measure (info score) <0.8 were discarded, and a threshold of 0.5 was applied on genotype uncertainty. Imputed variants with an MAF <1% were then excluded.

Discovery cohort association analysis

Univariate logistic regression analysis of non-genetic covariates (age, gender, BMI, CD4+ cell count) was undertaken for each hypersensitivity phenotype. Statistically significant variables ($P < 0.05$) were included in the subsequent logistic regressions to test for the association of each hypersensitivity phenotype with each SNP passing QC. All statistical analyses were undertaken using PLINK[27] and R.[31] Given prior associations between nevirapine hypersensitivity and *HLA* allele associations, it was felt reasonable to specify a Bonferroni-corrected *HLA*-wide significance threshold of $P < 2.5 \times 10^{-4}$, based on the presumption that there are usually <200 effective HLA allele tests.

Replication cohort genotyping, QC and association analysis

SNPs determined to have a nominally significant association with a nevirapine hypersensitivity phenotype ($P < 1 \times 10^{-5}$) in the discovery cohort were subsequently typed in the replication cohort using either the Sequenom MassArray iPLEX platform (Sequenom Inc., San Diego, CA, USA) or custom TaqMan real-time PCR SNP genotyping assays (Life Technologies, Paisley, UK) according to the manufacturer's protocols. SNPs were excluded if they failed to meet the genotype QC thresholds as outlined for the discovery cohort or if assay design software parameters prohibited their inclusion.

Logistic regression analysis of the replication cohort, including and excluding CD4+ count as a covariate, where appropriate (as determined in the discovery cohort), was carried out. Meta-analysis of combined discovery and replication cohorts was undertaken using a fixed-effects model with inverse-variant effect size weighting in GWAMA.[32]

Imputation of HLA allelotype and MHC locus

Imputation of *HLA-C* allelotype from the discovery cohort SNP array data was undertaken using HLA*IMP:02[33] (see Supplementary data).

HLA-C and ERAP gene–gene interactions

In cases and tolerant controls positive for carriage of the rs5010528 G allele, which was used as a proxy for *HLA-C*04:01*, we investigated both *ERAP1* (rs10050860 and rs30187) and *ERAP2* (rs2248374, rs2549782) SNPs

Table 1. Non-genetic data for nevirapine-tolerant and -hypersensitive patient cohorts included in both the main and replication analyses after sample exclusion based on genotyping QC criteria

	Cases [median (range)]					Controls [median (range)]					Phenotype (*n*)			
	n	age	female	BMI	CD4+	*n*	age	female	BMI	CD4+	NIR	HSS	SJS/TEN	DILI
Discovery cohort[a]	151	35 (17–69)	63%	20.5 (14.6–41.4)	235 (3–906)	182	35 (17–63)	60%	20.2 (13.7–36.4)	170 (2–677)	56	23	51	21
Replication cohort														
Malawi[b]	30	36 (24–58)	48%	19.8 (13.1–26.8)	207 (15–560)	29	36 (24–56)	47%	19.9 (14.9–30.3)	145 (25–314)	13	3	9[c]	6[c]
Uganda	32	37 (24–62)	73%	21.1 (8.9–39.6)	73 (17–201)	31	37 (25–53)	71%	21.3 (11.5–35.6)	66 (11–196)	14	8	2	10
Mozambique (SJS/TEN)[d]	27	31 (21–52)	100%	22.8 (12.4–35.0)	467 (107–906)	8	31 (23–41)	100%	27.5 (26.1–35.0)	556 (244–682)	–	–	27	–
Mozambique (DILI)	27	32 (23–43)	100%	23.7 (13.7–34.0)	324 (23–2305)	–	–	–	–	–	–	–	–	27
combined	116	34 (21–62)	78%	21.7 (8.9–39.6)	198 (15–2305)	68	36 (23–56)	64%	20.9 (11.5–35.6)	102 (11–682)	27	10	38	42

[a]CD4+ data missing for six cases.
[b]CD4+ data missing for four hypersensitive patients and BMI data missing for one hypersensitive patient.
[c]One patient presented with SJS/TEN and DILI phenotype.
[d]CD4+ data missing for three hypersensitive and four tolerant patients, and BMI data missing for six hypersensitive and four tolerant patients.

(using data from the Illumina array), which have previously been shown to interact with HLA-mediated immune diseases. Association of *ERAP1* and *ERAP2* SNPs with SJS/TEN risk was determined in the *HLA-C*04:01*-positive cohort (cases and controls) by logistic regression with CD4+ cell count as a covariate using PLINK.[27] A Bonferroni adjustment for multiple testing was applied with a significance threshold of $P = 0.125$.

Targeted sequencing of MHC region

Sixteen genomic DNA samples from nevirapine-induced SJS/TEN and 16 age- and gender-matched tolerant controls were carried forward for MHC-targeted sequencing. The methodology is detailed in the Supplementary data.

Allelotyping

HLA allelotyping was performed from raw FASTQ data files using Omixon Target v1.81 HLA Typing software and utilized the HLA database version 3.15.0. (Omixon Ltd, Budapest, Hungary).

In silico docking

In order to predict possible modes of interaction between nevirapine and *HLA-C*04:01*, *in silico* docking was undertaken. The methodology is detailed in the Supplementary data.

Results

Discovery cohort

A total of 333 samples (151 cases and 182 controls) out of 352 passed QC. Of the 19 excluded samples, 9 failed heterozygosity checks (outliers by >3 SD), 8 failed identity checks and 2 failed the call rate threshold. Multi-dimensional analysis for population stratification (Figure S1) demonstrated no population outliers. In total, 817 728 SNPs passed QC and were carried over for imputation with the 1000 genomes panel. Imputation produced a dataset of 14218511 variants. Cohort characteristics are shown in Table 1. We considered five nevirapine-induced hypersensitivity phenotypes for analysis—NIR, HSS, SJS/TEN, DILI (Table 1)—and also combined these different phenotypes into an overall hypersensitivity group.

Univariate logistic regression analysis showed CD4+ cell count to be a statistically significant variable for NIR ($P = 0.016$), SJS ($P = 0.003$) and all hypersensitivity cases ($P = 2.5 \times 10^{-6}$). Therefore, we included CD4+ cell count as a covariate in the SNP logistic regression model for these three phenotypes. Multidimensional scaling (MDS) variables were not included as covariates in the logistic regression since the population stratification analysis suggested that the cohort was homogeneous and genomic control was unnecessary (Figure S1). From the genome-wide logistic regression analyses, we identified 15 SNPs with $P < 1 \times 10^{-5}$, with at least one of the five different phenotypes analysed (Figure 2; summarized in Table 2). No variant reached genome-wide significance.

Replication cohort

Of the 15 SNPs considered for replication, one (rs150223496) could not be typed due to proximal sequence constraints of the Sequenom assay design process, and QC failure for TaqMan genotyping Hardy-Weinberg equilibrium (HWE)

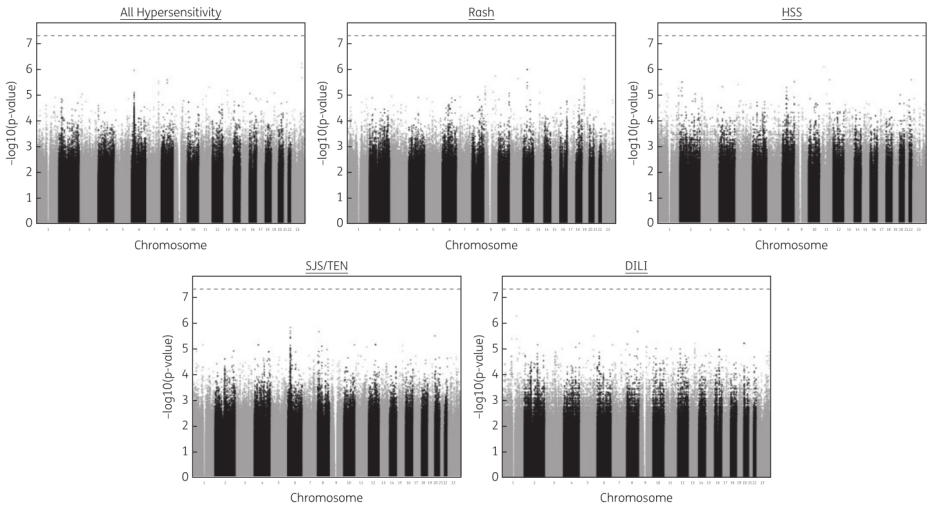


Figure 2. Manhattan plots of association for logistic regression SNP analysis for the five defined nevirapine hypersensitivity phenotypes. The phenotypes, all hypersensitivity, NIR and SJS/TEN show *P* values from logistic regression incorporating CD4+ count as a covariate. For HSS and DILI no covariates were incorporated. The broken line indicates genome-wide significance ($P = 5 \times 10^{-8}$).

Table 2. Top SNPs identified in association with a nevirapine hypersensitivity phenotype within the main cohort analysis

								Logistic regression	
Phenotype	SNP	Chr	Position (GRCh37 .p13)	Reference allele	Associated allele	Gene	Typed/ imputed	*P* value	OR (95% CI)
All hypersensitivity	rs34213790	6	31318579	G	A	3′ of *HLA-B*	imputed	1.15×10^{-6}	2.34 (1.66–3.30)
($n = 151$)	rs11764223	7	137742056	T	G	*AKR1D1*	imputed	2.99×10^{-6}	0.45 (0.33–0.63)
	rs11988543	8	68596362	T	A	*CPA6*	imputed	2.69×10^{-6}	3.37 (2.03–5.60)
	rs9527426	13	56208216	T	C	*MIR5007*	typed	9.86×10^{-6}	2.25 (1.58–3.21)
	rs35990155	23	11827017	G	T	*KIAA1210*	imputed	6.33×10^{-7}	5.65 (2.86–11.17)
Rash ($n = 56$)	rs10815440	9	661057	G	A	*KANK1*	typed	8.86×10^{-6}	5.80 (2.68–12.67)
	rs115848367	10	128045499	A	C	*ADAM12*	imputed	8.83×10^{-7}	4.70 (2.38–9.30)
	rs74373347	12	48919428	A	G	*OR8S1*	imputed	1.08×10^{-6}	11.05 (4.12–29.04)
	rs6511720	19	11202306	G	T	*LDLR*	typed	5.49×10^{-6}	4.66 (2.40–9.05)
SJS/TEN ($n = 51$)	rs5010528	6	31241032	A	G	*HLA-C*	typed	4.13×10^{-6}	4.75 (2.45–9.22)
	rs150223496	15	76421885	G	A	*C15orf27*	imputed	8.20×10^{-6}	17.34 (4.95–60.74)
DILI ($n = 21$)	rs147773805	1	107741393	A	T	*NTNG1*	imputed	4.47×10^{-6}	11.05 (3.96–30.83)
	rs114693001	9	76074363	A	T	intergenic	imputed	8.82×10^{-6}	9.35 (3.49–25.06)
	rs142213069	13	77962182	G	C	5′ of *MYCBP2*	imputed	6.64×10^{-6}	14.42 (4.51–46.04)
	rs6139258	20	3958616	T	C	*RNF24*	typed	6.64×10^{-6}	14.42 (4.52–46.04)

P values and ORs were determined by logistic regression with CD4+ cell count as a covariate (except for HSS and DILI).

Table 3. Logistic regression analysis of candidate SNPs in the nevirapine hypersensitivity replication cohort

Phenotype	SNP	Chr	Position (GRCh37 .p13)	Genotyping platform	Reference allele	Associated allele	Gene	Logistic regression	
								P value	OR (95% CI)
All hypersensitivity ($n = 62$)	rs34213790	6	31318579	iPLEX	G	A	3′ of *HLA-B*	0.10	1.56 (0.94–2.60)
	[a]rs12112517	7	137743167	iPLEX	A	C	*AKR1D1*	0.21	1.42 (0.82–2.46)
	rs11988543	8	68596362	TaqMan	T	A	*CPA6*	0.77	1.13 (0.54–2.28)
	rs9527426	13	56208216	iPLEX	T	C	*MIR5007*	0.83	1.06 (0.62–1.80)
	rs35990155	23	11827017	iPLEX	G	T	*KIAA1210*	0.86	0.94 (0.48–1.86)
Rash ($n = 27$)	rs10815440	9	661057	iPLEX	G	A	*KANK1*	0.99	3.3×10^{-10} (0–∞)
	rs115848367	10	128045499	iPLEX	A	C	*ADAM12*	0.70	1.18 (0.51–2.70)
	rs74373347	12	48919428	iPLEX	A	G	*OR8S1*	0.76	0.81 (0.21–3.11)
	rs6511720	19	11202306	iPLEX	G	T	*LDLR*	0.86	1.07 (0.51–2.24)
SJS/TEN ($n = 38$)	rs5010528	6	31241032	TaqMan	A	G	*HLA-C*	**0.02**	5.33 (1.37–20.80)
	[b]rs150223496	15	76421885	NA	G	A	*C15orf27*	NA	NA
DILI ($n = 42$)	rs1730858	1	107619244	iPLEX	G	A	*NTNG1*	0.61	1.57 (0.27–8.97)
	[c]rs114693001	9	76074363	TaqMan	A	T	intergenic	0.99	2.6×10^{-9} (0–∞)
	[c]rs142213069	13	77962182	iPLEX	G	C	5′ of *MYCBP2*	0.99	2.4×10^{-9} (0–∞)
	rs6139258	20	3958616	iPLEX	T	C	*RNF24*	**0.03**	11.17 (1.29–96.38)

SNP associations below arbitrary statistical significance ($P < 0.1$) are highlighted in bold. NA denotes statistical analysis not applicable.
[a]SNPs within the association signal that are substituted from the discovery cohort SNP (high LD) due to genotyping assay design constraints.
[b]SNP signal where replication in the replication cohort was not possible.
[c]SNPs that were not be typed in the Mozambique DILI cohort ($n = 15$ cases). Mozambique individuals were also omitted from the 'all hypersensitivity' analysis.

Table 4. Meta-analysis of significantly associated SNPs in the discovery and replication cohorts

SNP	Gene	Phenotype	Cohort	Case/ control	*n*	MAF	Logistic regression		Meta-analysis	
							P	OR (95% CI)	*P*	OR (95% CI)
rs34213790	3′ of *HLA-B*	all hypersensitivity	discovery	case	151	0.45	1.15×10^{-6}	2.34 (1.66–3.30)	–	–
				control	182	0.37				
			replication	case	27	0.50	0.10	1.56 (0.94–2.60)	6.69×10^{-7}	2.06 (1.55–2.74)
				control	60	0.38				
rs5010528	*HLA-C*	SJS/TEN	discovery	case	51	0.36	4.13×10^{-6}	4.75 (2.45–9.22)	–	–
				control	182	0.14				
			replication	case	38	0.38	6.03×10^{-3}	5.12 (1.60–16.41)	8.47×10^{-8}	4.84 (2.71–8.61)
				control	68	0.12				
rs6139258	*RNF24*	DILI	discovery	case	21	0.19	6.64×10^{-6}	14.42 (4.52–46.04)	–	–
				control	182	0.02				
			replication	case	42	0.07	0.03	11.17 (1.29–96.38)	5.66×10^{-7}	13.62 (4.90–37.84)
				control	68	0.01				

Data shown are for analysis undertaken with covariates (CD4+ cell count) included in the regression model (except DILI) for both the discovery and replication cohorts (± supplementary cohort for SJS/TEN and DILI).

P value >0.0001, call rate >90%. Thus, 14 SNPs were carried forward for analysis. A total of 59 Malawian samples (30 cases and 29 controls) and 63 Ugandan samples (32 cases and 31 controls) passed QC (call rate >90%) as described in Table 1. Due to sample constraints, SNP signals identified in the discovery cohort for the 'all hypersensitivity' phenotype were not typed in the samples from Mozambique in the replication cohort.

For nevirapine-induced DILI (Table 3), combining the discovery and replication cohorts for the SNP rs6139258 in the *RNF24* locus strengthened the association [$P = 5.7 \times 10^{-7}$, OR 13.62 (95% CI 4.90–37.84)] (Table 4). A single SNP (rs5010528 in the *HLA-C* locus) showed an association with SJS/TEN in the replication cohort (38 cases, 59 controls) [$P = 0.006$; OR 5.12 (95% CI 1.60–16.42)]. Combining the discovery and replication cohorts strengthened the

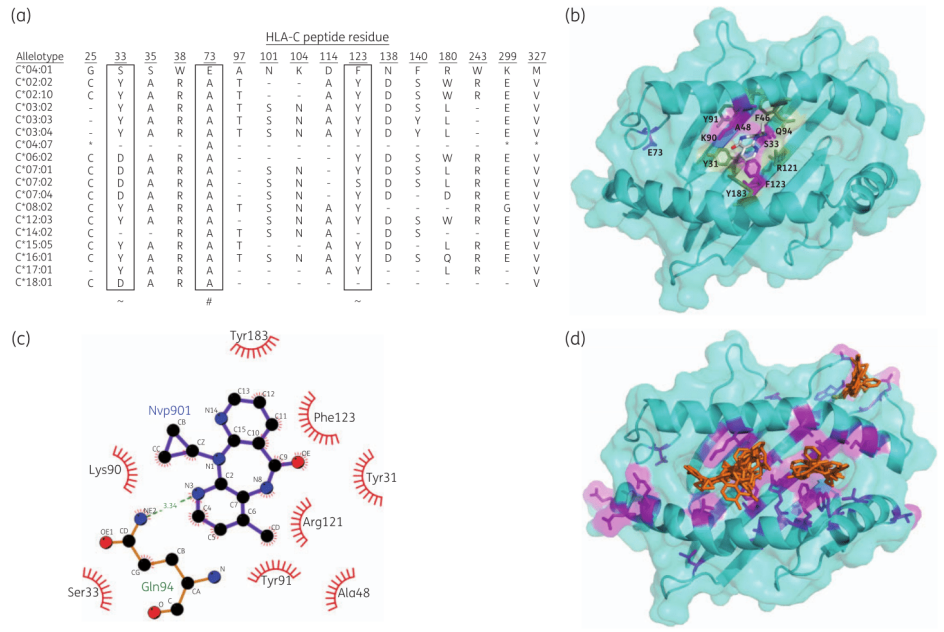


Figure 3. *In silico* docking of HLA-C*04:01 and nevirapine. (a) Specific peptide residues present in the different HLA-C allelotypes previously identified in the Malawian cohort ($n = 116$). – shows continuity with the reference peptide (C*04:01) and * identifies peptides not sequenced in an allelotypes. ~ denotes putative nevirapine interaction and # denotes residue substituted by SNP (rs1050409) identified in GWAS. Residue reference numbers are as defined by the *in silico* model. (b) HLA-C*04:01/nevirapine docking mode of conformation with the highest predicted affinity (lowest score) as produced using the PyMOL software. Key interacting residues are highlighted. (c) A LIGPLOT+[40] 2D schematic representation of the interaction of nevirapine with HLA-C*04:01 PBD residues in the highest affinity predicted docking conformation mode. The broken line indicates a hydrogen bond and radiating lines indicate hydrophobic interactions. (d) HLA-C*04:01/nevirapine (orange) docking mode of conformation for the 20 highest predicted affinity (lowest scores) as produced using the PyMOL software.

overall association, which approached genome-wide significance [$P = 8.5 \times 10^{-8}$, OR 4.84 (95% CI 2.71–8.61)]. No positive signals were identified for the 'all hypersensitivity' phenotype (62 cases, 59 controls).

HLA-C allelotype imputation

Overall allelotype imputation from the SNP array data using HLA*IMP demonstrated 71.5% concordance with the HLA typing obtained for 116 of our patients using the sequence-based methodology. However the ability of the imputation to correctly call *HLA-C*04:01* alleles was 90%.

Within the 116 patients for which HLA allelotyping and SNP array genotype data were available, *HLA-C-04:01* allele carriage co-occurred with the rs5010528 G allele in 112/116 cases (96.5%). For the imputed *HLA* allelotype data, *C*04:01* co-occurred with rs5010528 G in 303/333 cases (91%).

The initial discovery logistic regression analysis demonstrated two non-synonymous SNPs in the *HLA-C* locus associated with SJS/TEN that were in absolute linkage disequilibrium (LD) with rs5010528 (Table 5). The first SNP (rs146911342) encodes a valine-to-methionine amino acid substitution at residue 327 and the second (rs1050409) encodes an alanine to glutamic acid at residue 73 (close to the peptide binding domain), which was also associated with SJS/TEN [$P = 4.1 \times 10^{-6}$, OR 4.75 (95% CI2.45–9.23)]. Both are key defining residues of the *HLA-C*04* allelotypes as defined in the HLA-IMGT database[34] and within the Malawian cohort (Figure 3). Verification of either SNP in our discovery or replication cohorts via other genotyping methodologies was not possible due to sequence constraints in assay design (the proximal nucleotide sequence for primer design was not sufficiently unique or contained a restrictive number of other genetic variants).

However, targeted sequencing of the HLA locus in 16 SJS/TEN and 16 tolerant controls (Table 5) suggested that both the

Table 5. Association of nevirapine-induced SJS/TEN and imputed SNPs of the *HLA-C* locus from the main cohort and the targeted sequencing cohort

					GWAS (discovery cohort)					Targeted sequencing cohort				
SNP	bp (GRCh37.p13)	A_1/A_2	Amino acid substitution	typed/imputed	SJS/TEN MAF (n=51)	control MAF (n=182)	P	OR (95% CI)	SJS/TEN MAF (n = 16)	control MAF (n = 16)	P	OR (95% CI)	LD with rs5010528 (D')
rs146911342	31 237 779	C/T	p.V327M	imputed	0.36	0.15	4.13×10^{-6}	4.75 (2.45–9.23)	0.34	0.13	0.013	34.14 (2.07–561.6)	1
rs41562714	31 238 538	G/C		imputed	0.36	0.15	4.13×10^{-6}	4.75 (2.45–9.23)	0.34	0.13	0.013	34.14 (2.07–561.6)	1
rs10504409	31 239 501	G/T	p.A73E	imputed	0.36	0.15	4.13×10^{-6}	4.75 (2.45–9.23)	0.34	0.13	0.013	34.14 (2.07–561.6)	1
rs41553018	31 239 742	G/C		imputed	0.36	0.15	3.71×10^{-6}	4.75 (2.45–9.23)	0.34	0.13	0.013	34.14 (2.07–561.6)	1
rs4361609	31 240 635	G/C		imputed	0.36	0.15	4.13×10^{-6}	4.75 (2.45–9.23)	0.34	0.13	0.013	34.14 (2.07–561.6)	1
rs5010528	31 241 032	A/G		typed	0.36	0.15	4.13×10^{-6}	4.75 (2.45–9.23)	0.34	0.13	0.013	34.14 (2.07–561.6)	–
rs58019823	31 241 215	A/G		imputed	0.36	0.15	4.13×10^{-6}	4.75 (2.45–9.23)	–	–	–	–	NA
[a]rs2524087	31241 294	C/G		imputed	0.37	0.15	3.71×10^{-6}	4.73 (2.45–9.14)	–	–	–	–	NA
[a]rs59103503	31 287 944	G/T		imputed	0.37	0.15	2.75×10^{-6}	5.00 (2.55–9.80)	–	–	–	–	NA
[a]rs77641320	31 298 229	C/G		imputed	0.35	0.14	1.64×10^{-6}	5.53 (2.75–11.12)	–	–	–	–	NA
[a]rs9468965	31 300 247	T/A		imputed	0.44	0.19	2.24×10^{-6}	4.61 (2.24–8.69)	–	–	–	–	NA
[a]rs35364987	31 309 423	T/C		imputed	0.49	0.27	8.01×10^{-6}	3.76 (2.10–6.74)	–	–	–	–	NA
[a]rs35435945	31 311 374	T/G		imputed	0.49	0.27	9.83×10^{-6}	3.72 (2.08–6.67)	–	–	–	–	NA
[a]rs35278939	31 319 780	G/A		imputed	0.47	0.23	8.01×10^{-6}	4.04 (2.19–7.47)	–	–	–	–	NA

The list comprises all SNPs with a P value of $<1 \times 10^{-5}$ in the main cohort analysis. Statistical significance and OR (95% CI) for the targeted sequencing cohort is determined by logistic regression with CD4+ cell count as covariate. NA denotes LD indeterminable in targeted sequencing cohort as SNP typing not available.
[a]SNP was imputed in the discovery cohort analysis but not detected and or called in the sequencing data.

imputed non-synonymous SNPs may be in absolute LD with the original discovery cohort signal SNP (rs5010528), again demonstrating a significant association with nevirapine-induced SJS/TEN. Allelotype inference from the targeted sequencing SNP data also confirmed that both non-synonymous SNPs were in 100% co-occurrence with *HLA-C*04:01* (data not shown). Our data do not show an association between any of the other *HLA* gene loci and other nevirapine-induced hypersensitivity phenotypes.

*HLA-C*04:01 and ERAP1 and ERAP2 SNP interactions*

Given the previously reported interactions between ERAP genes and HLA class I-mediated diseases, in particular ankylosing spondylitis,[35] we determined whether there was an interaction with the carriage of *HLA-C*04:01* using the rs5010528 G allele as a proxy SNP (Table S1). There was no significant association ($P > 0.05$) between the *ERAP1* variants and SJS/TEN risk in carriers of *HLA-C*04:01*. However, both *ERAP2* variants showed a nominal association with SJS/TEN risk [$P = 0.019$, OR 0.43 (95% CI 0.21–0.87)], though this did not pass the Bonferroni threshold for multiple testing ($P = 0.0125$).

In silico docking

In light of the association observed between nevirapine-induced SJS/TEN and an SNP (rs1050409) encoding an amino acid substitution at residue 73 of HLA-C (p.A73E), *in silico* docking was undertaken to determine the possible effect of the residue substitution on nevirapine binding. The data suggest that none of the predicted modes of nevirapine docking conformation interact with residue 73, which appears to be on the periphery of the peptide-binding domain (Figure 3). The lowest scoring (predicted highest affinity) mode highlights an interaction between nevirapine and residues 33 and 123 in the B pocket (Figure 3). In *HLA-C*04:01*, residues 33 and 123 are serine and phenylalanine respectively (Figure 3). The majority of other allelotypes do not possess these particular residues (with the exception of C*04:07 and C*14:02). Docking of the metabolite 12-hydroxy-nevirapine was also undertaken, since it has also been suggested as potentially responsible for nevirapine-induced adverse drug reactions;[36] these were in general agreement with those for nevirapine, in that docking seems to take place around the B pocket (e.g. near residues 33 and 123), but with more variability in the different modes predicted than for nevirapine. None of the predicted modes interacted with residue 73 (data not shown). Taken together, the docking results suggest that binding of either nevirapine or 12-hydroxy-nevirapine around the centre of the peptide-binding regions is likely to be important in the mechanism of the immune-mediated reaction.

Discussion

The investigation of genetic factors predisposing to serious adverse drug reactions is challenging because of their rarity. Despite this, we have assembled one of the largest cohorts of patients with clinically well-characterized nevirapine hypersensitivity, including SJS/TEN. GWAS analysis of our Malawian discovery cohort ($n = 333$) identified 15 polymorphisms having a suggestive association with nevirapine hypersensitivity (Table 2). Subsequent analysis of these variants in our replication cohort suggested that three of the SNPs may be potential risk factors (Table 3): rs34213790 3′ of the *HLA-B* gene locus with all hypersensitivity phenotypes; rs5010528 in the *HLA-C* gene locus with SJS/TEN, and rs6139258 in the *RNF24* gene locus with DILI. The weakest of the above three association signals, SNP rs34213790, is unlikely to be an independent marker of nevirapine hypersensitivity in general, and its association may be due to a haplotype effect between *HLA-C*04:01* (rs5010528; see below) and B allelotypes.

SNP rs6139258 in the *RNF24* gene locus only marginally failed to pass the Bonferroni threshold of significance in the replication cohort. Very little is known regarding the function of *RNF24*. However, it is known that it is a protein that interacts with transient receptor potential cation channel 6 (TRPC6),[37] a receptor-activated channel, expressed in liver cells,[38] which plays a role in cellular calcium homeostasis. *TRPC6* has been suggested to play a role in hepatoma cell-line proliferation, possibly via a cyclin D-modulated mechanism.[38] Thus *RNF24* may have some biological plausibility in the pathogenesis of nevirapine-induced liver injury, and merits further investigation in additional patients with nevirapine-induced DILI, and functional work to uncover the possible mechanisms (if any) of the association.

The most compelling of the three signals, rs5010528, gave an OR of 4.84 for nevirapine-induced SJS/TEN, and was replicated in patients from three countries (Malawi, Uganda and Mozambique) at the Bonferroni threshold ($P < 0.05$), approaching genome-wide significance in the combined analysis ($P = 8.5 \times 10^{-8}$) (Table 4). SNP rs5010528 is located within the *HLA-C* gene locus. High co-occurrence of rs5010528 with *HLA-C*04:01* was observed (96.5%) in 116 patients within this study who had previously been HLA typed by sequence-based methods. The association between nevirapine and rs5010528 (as a proxy for *C*04:01*) can be considered statistically significant when applying an HLA-wide significance threshold of $P < 2.5 \times 10^{-4}$. Additionally, *HLA-C* allelotypes imputed from the SNP array data also showed a high co-occurrence (91%) in the main study cohort, suggesting rs5010528 may be a good proxy for *HLA-C*04:01*. Thus, the GWAS data appear to confirm our previous finding[21] associating *HLA-C*04:01* with nevirapine-induced SJS/TEN. Of note, the association of rs5010528 with other hypersensitivity phenotypes was not as strong, suggesting that the risk conferred by rs5010528 and thus *HLA-C*04:01* is specific for nevirapine-induced SJS/TEN. The reason for this is unclear, and will require further investigation.

In terms of clinical utility, rs5010528 appears to have little potential as a pre-emptive genetic test. Indeed, based on a prevalence of SJS/TEN in our prospective cohort of 1.07% and assuming a dominant mode of inheritance, the positive (PPV) and negative (NPV) predictive values were 2.8% and 42.4% respectively. For the *RNF24* variant (rs6139258) the PPV, based on a prevalence of DILI of 0.63%, is also very low (0.2%).

Only one previous GWAS investigating nevirapine hypersensitivity has been reported, but in a smaller Thai population (72 cases, 77 controls).[39] Patients had a wide variety of rashes, with only 11 grade 4 cases (6.9%), which would be equivalent to our cases with SJS/TEN. The SNP rs9461684 in the *HLA-C* locus was significantly associated with nevirapine rash, but no *HLA* allelotype imputation or *HLA* sequencing was carried out. In our data, rs9461684 is in high LD with our top SNP, rs5010528 ($D' = 1.0$, $r^2 = 0.972$). The discrepancy may be a result of the different LD patterns in the

different ethnic groups studied, as well as the much lower numbers of patients with serious skin reactions in the Thai study.

From the imputed SNP data of the discovery cohort and targeted resequencing data (Table 5), it is clear that rs5010528 is in LD with a functional non-synonymous SNP (rs1050409) that leads to an alanine-to-glutamic acid substitution at residue 73, which lies near to the peptide-binding domain of the HLA-C protein. However, *in silico* modelling suggested that this residue does not interact with nevirapine in any docking conformation. Two other residues of *HLA-C*04:01* (33 and 123 in the model) appear to be the key interactors in the majority of the predicted modes of nevirapine docking (Figure 3), including the most favoured. However, it should be noted that this is a predictive model and further analysis of the HLA-C/nevirapine complex is needed to further elucidate the potential for docking. The association signal at residue 73 (rs1050409) is likely to be a proxy for the 33 and 123 residues also present in HLA-C*04:01. However, this work has provided the first evidence that nevirapine binds to the B pocket of *HLA-C*04:01*.

ERAP gene variants interact in a protective manner in HLA-mediated diseases such as ankylosing spondylitis in individuals who carry the risk *HLA* alleles.[35] ERAP1 and ERAP2 are enzymes involved in antigenic peptide precursor trimming prior to loading into HLA class I molecules (and may thus alter the peptidome) and may potentially also alter the expression of the risk HLA class I allele. To our knowledge, this is one of the first examinations of whether there is interaction between drug-induced HLA disease and the *ERAP* genes. We were, however, unable to detect an interaction between *ERAP1* variants and *HLA-C*04:01* in African patients with SJS/TEN. However, a nominal association ($P = 0.019$) was observed for both *ERAP2* SNPs (Table S1). A limitation of our analysis is the small sample size, particularly given the much larger numbers that have been studied in ankylosing spondylitis. Nevertheless, the possibility of an association with ERAP2 is intriguing, and needs further investigation not only with nevirapine-induced hypersensitivity, but also with other HLA-related adverse drug reactions.

In identifying an SNP in the *HLA-C* locus that appears to be a proxy for the *HLA-C*04:01* allele, as a risk factor for nevirapine-induced SJS/TEN, this study has added further weight to existing evidence. The data generated also suggest that, in sub-Saharan African HIV patients, no other strong, significant genetic risk factors for nevirapine hypersensitivity exist that could be utilized as clinical predictive markers. However, the data are valuable in terms of the mechanistic insights they provide. Additionally, *in silico* analysis has identified two putative HLA-C peptide residues that are predicted to be key for the binding of nevirapine, which warrant further investigation as to their role in the pathogenesis of SJS/TEN. Further work is also needed to determine the reasons for organ-specific toxicities in different patients.

Acknowledgements

We thank the patients and staff of the ART clinic of Queen Elizabeth Central Hospital, Blantyre, in particular Mr S. Kaunda, Clinical Officer. We also thank Dr Christiane Hertz-Fowler, Dr Margaret Hughes, Dr Lisa Olohan and Dr Anita Lucaci from the Centre for Genomic Research for undertaking the library preparation and sequencing of the MHC region in the cohort of 32 DNA samples.

Funding

The initial GWAS was funded by the International Serious Adverse Events Consortium (iSAEC). The iSAEC is a non-profit organization dedicated to identifying and validating DNA variants useful in predicting the risk of drug-related serious adverse events. The Consortium brings together the pharmaceutical industry, regulatory authorities and academic centres to address clinical and scientific issues associated with the genetics of drug-related serious adverse events. The iSAEC's current funding members include: Abbott, Amgen, AstraZeneca, Daiichi Sankyo, GlaxoSmithKline, Merck, Novartis, Pfizer, Takeda and the Wellcome Trust. Mas Chaponda was funded by a 3 year Wellcome Trust training fellowship WT078857MA administered through the University of Liverpool. Malawi-Liverpool-Wellcome Trust Clinical Research Programme is funded through a Core Programme Grant award from the Wellcome Trust. Munir Pirmohamed is a National Institute for Health Research Senior Investigator, and also wishes to thank the MRC Centre for Drug Safety Science for support.

The DART study was supported by the UK Medical Research Council (grant number G0600344), the UK Department for International Development and the Rockefeller Foundation.

Andrew P. Morris is a Wellcome Trust Senior Research Fellow in Basic Biomedical Science (grant number WT098017).

Louise Y. Takeshita is funded by a PhD fellowship from CNPq (National Council for Scientific and Technological Development, Brazil).

Panos Deloukas' work forms part of the research themes contributing to the translational research portfolio of Barts Cardiovascular Biomedical Research Unit which is supported and funded by the National Institute for Health Research.

Transparency declarations

All authors: no conflicts of interest to declare.

GlaxoSmithKline, Gilead Sciences, Boehringer-Ingelheim and AbbVie donated drugs for the DART study.

Supplementary data

Figure S1 and Table S1 are available as Supplementary data at *JAC* Online.

References

1 van Leth F, Phanuphak P, Ruxrungtham K *et al*. Comparison of first-line antiretroviral therapy with regimens including nevirapine, efavirenz, or both drugs, plus stavudine and lamivudine: a randomised open-label trial, the 2NN Study. *Lancet* 2004; **363**: 1253–63.

2 Siegfried NL, Van Deventer PJ, Mahomed FA *et al*. Stavudine, lamivudine and nevirapine combination therapy for treatment of HIV infection and AIDS in adults. *Cochrane Database Syst Rev* 2006; **19**: issue CD004535.

3 Phillips E, Gutierrez S, Jahnke N *et al*. Determinants of nevirapine hypersensitivity and its effect on the association between hepatitis C status and mortality in antiretroviral drug-naive HIV-positive patients. *AIDS* 2007; **21**: 1561–8.

4 Wit FW, Kesselring AM, Gras L *et al*. Discontinuation of nevirapine because of hypersensitivity reactions in patients with prior treatment experience, compared with treatment-naive patients: the ATHENA cohort study. *Clin Infect Dis* 2008; **46**: 933–40.

5 Fagot JP, Mockenhaupt M, Bouwes-Bavinck JN *et al*. Nevirapine and the risk of Stevens-Johnson syndrome or toxic epidermal necrolysis. *AIDS* 2001; **15**: 1843–8.

6 Mittmann N, Knowles SR, Koo M *et al*. Incidence of toxic epidermal necrolysis and Stevens-Johnson syndrome in an HIV cohort: an observational, retrospective case series study. *Am J Clin Dermatol* 2012; **13**: 49–54.

7 De Maat MM, Mathot RA, Veldkamp AI *et al*. Hepatotoxicity following nevirapine-containing regimens in HIV-1-infected individuals. *Pharmacol Res* 2002; **46**: 295–300.

8 Riska P, Lamson M, MacGregor T *et al*. Disposition and biotransformation of the antiretroviral drug nevirapine in humans. *Drug Metab Dispos* 1999; **27**: 895–901.

9 Jinno H, Tanaka-Kagawa T, Ohno A *et al*. Functional characterization of cytochrome P450 2B6 allelic variants. *Drug Metab Dispos* 2003; **31**: 398–403.

10 Lang T, Klein K, Fischer J *et al*. Extensive genetic polymorphism in the human CYP2B6 gene with impact on expression and function in human liver. *Pharmacogenetics* 2001; **11**: 399–415.

11 Rotger M, Colombo S, Furrer H *et al*. Influence of CYP2B6 polymorphism on plasma and intracellular concentrations and toxicity of efavirenz and nevirapine in HIV-infected patients. *Pharmacogenet Genomics* 2005; **15**: 1–5.

12 Penzak SR, Kabuye G, Mugyenyi P *et al*. Cytochrome P450 2B6 (CYP2B6) G516T influences nevirapine plasma concentrations in HIV-infected patients in Uganda. *HIV Med* 2007; **8**: 86–91.

13 Yuan J, Guo S, Hall D *et al*. Toxicogenomics of nevirapine-associated cutaneous and hepatic adverse events among populations of African, Asian, and European descent. *AIDS* 2011; **25**: 1271–80.

14 Haas DW, Bartlett JA, Andersen JW *et al*. Pharmacogenetics of nevirapine-associated hepatotoxicity: an Adult AIDS Clinical Trials Group collaboration. *Clin Infect Dis* 2006; **43**: 783.

15 Martin AM, Nolan D, James I *et al*. Predisposition to nevirapine hypersensitivity associated with HLA-DRB1*0101 and abrogated by low CD4 T-cell counts. *AIDS* 2005; **19**: 97–9.

16 Vitezica ZG, Milpied B, Lonjou C *et al*. HLA-DRB1*01 associated with cutaneous hypersensitivity induced by nevirapine and efavirenz. *AIDS* 2008; **22**: 540–1.

17 Likanonsakul S, Rattanatham T, Feangvad S *et al*. HLA-Cw*04 allele associated with nevirapine-induced rash in HIV-infected Thai patients. *AIDS Res Ther* 2009; **6**: 22.

18 Gao S, Gui XE, Liang K *et al*. HLA-dependent hypersensitivity reaction to nevirapine in Chinese Han HIV-infected patients. *AIDS Res Hum Retroviruses* 2011; **28**: 540–3.

19 Gatanaga H, Yazaki H, Tanuma J *et al*. HLA-Cw8 primarily associated with hypersensitivity to nevirapine. *AIDS* 2007; **21**: 264–5.

20 Chantarangsu S, Mushiroda T, Mahasirimongkol S *et al*. HLA-B*3505 allele is a strong predictor for nevirapine-induced skin adverse drug reactions in HIV-infected Thai patients. *Pharmacogenet Genomics* 2009; **19**: 139–46.

21 Carr DF, Chaponda M, Jorgensen AL *et al*. Association of human leukocyte antigen alleles and nevirapine hypersensitivity in a Malawian HIV-infected population. *Clin Infect Dis* 2013; **56**: 1330.

22 Kenna TJ, Robinson PC, Haroon N. Endoplasmic reticulum aminopeptidases in the pathogenesis of ankylosing spondylitis. *Rheumatology (Oxford)* 2015; **54**: 1549–56.

23 Mugyenyi P, Walker AS, Hakim J *et al*. Routine versus clinically driven laboratory monitoring of HIV antiretroviral therapy in Africa (DART): a randomised non-inferiority trial. *Lancet* 2010; **375**: 123–31.

24 Ciccacci C, Borgiani P, Ceffa S *et al*. Nevirapine-induced hepatotoxicity and pharmacogenetics: a retrospective study in a population from Mozambique. *Pharmacogenomics* 2010; **11**: 23–31.

25 Ciccacci C, Di Fusco D, Marazzi MC *et al*. Association between CYP2B6 polymorphisms and nevirapine-induced SJS/TEN: a pharmacogenetics study. *Eur J Clin Pharmacol* 2013; **69**: 1909–16.

26 Munderi P, Snowden WB, Walker AS *et al*. Distribution of HLA-B alleles in a Ugandan HIV-infected adult population: NORA pharmacogenetic substudy of DART. *Trop Med Int Health* 2011; **16**: 200–4.

27 Purcell S, Neale B, Todd-Brown K *et al*. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007; **81**: 559–75.

28 Delaneau O, Zagury JF, Marchini J. Improved whole-chromosome phasing for disease and population genetic studies. *Nat Methods* 2013; **10**: 5–6.

29 Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* 2009; **5**: e1000529.

30 Marchini J, Howie B. Genotype imputation for genome-wide association studies. *Nat Rev Genet* 2010; **11**: 499–511.

31 R Core Team (2013). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0. http://www.R-project.org.

32 Magi R, Morris AP. GWAMA: software for genome-wide association meta-analysis. *BMC Bioinformatics* 2010; **11**: 288.

33 Dilthey A, Leslie S, Moutsianas L *et al*. Multi-population classical HLA type imputation. *PLoS Comput Biol* 2013; **9**: e1002877.

34 Robinson J, Halliwell JA, Hayhurst JD *et al*. The IPD and IMGT/HLA database: allele variant databases. *Nucleic Acids Res* 2015; **43**: D423–31.

35 Evans DM, Spencer CC, Pointon JJ *et al*. Interaction between ERAP1 and HLA-B27 in ankylosing spondylitis implicates peptide handling in the mechanism for HLA-B27 in disease susceptibility. *Nat Genet* 2011; **43**: 761–7.

36 Sharma AM, Novalen M, Tanino T *et al*. 12-OH-nevirapine sulfate, formed in the skin, is responsible for nevirapine-induced skin rash. *Chem Res Toxicol* 2013; **26**: 817–27.

37 Lussier MP, Lepage PK, Bousquet SM *et al*. RNF24, a new TRPC interacting protein, causes the intracellular retention of TRPC. *Cell Calcium* 2008; **43**: 432–43.

38 El Boustany C, Bidaux G, Enfissi A *et al*. Capacitative calcium entry and transient receptor potential canonical 6 expression control human hepatoma cell proliferation. *Hepatology* 2008; **47**: 2068–77.

39 Chantarangsu S, Mushiroda T, Mahasirimongkol S *et al*. Genome-wide association study identifies variations in 6p21.3 associated with nevirapine-induced rash. *Clin Infect Dis* 2011; **53**: 341–8.

40 Laskowski RA, Swindells MB. LigPlot+: multiple ligand-protein interaction diagrams for drug discovery. *J Chem Inf Model* 2011; **51**: 2778–86.

Appendix B

The following table contains a subset of the KIR and disease associations used in the analysis presented in Chapter 2.

PMID	Country	KIR	Association	Disease Class	Disease Name
24185760	Canada	2DL2	Susceptibility	Autoimmune	Arthritis, Psoriatic
24185760	Canada	2DS2	Susceptibility	Autoimmune	Arthritis, Psoriatic
17882223	Poland	2DL1	Susceptibility	Autoimmune	Arthritis, Rheumatoid
26658904	Iran	2DL1	Susceptibility	Autoimmune	Arthritis, Rheumatoid
26658904	Iran	2DL2	Protection	Autoimmune	Arthritis, Rheumatoid
24912006	India	2DL2	Protection	Autoimmune	Arthritis, Rheumatoid
16641046	Taiwan	2DL2	Susceptibility	Autoimmune	Arthritis, Rheumatoid
22960345	Mexico	2DL2	Susceptibility	Autoimmune	Arthritis, Rheumatoid
27251940	Mexico	2DL2	Susceptibility	Autoimmune	Arthritis, Rheumatoid
22960345	Mexico	2DL3	Protection	Autoimmune	Arthritis, Rheumatoid
24912006	India	2DL3	Protection	Autoimmune	Arthritis, Rheumatoid
17882223	Poland	2DL3	Susceptibility	Autoimmune	Arthritis, Rheumatoid
26658904	Iran	2DL5	Protection	Autoimmune	Arthritis, Rheumatoid
24912006	India	2DP1	Protection	Autoimmune	Arthritis, Rheumatoid
26658904	Iran	2DP1	Susceptibility	Autoimmune	Arthritis, Rheumatoid
17882223	Poland	2DS2	Protection	Autoimmune	Arthritis, Rheumatoid
26658904	Iran	2DS2	Protection	Autoimmune	Arthritis, Rheumatoid
16641046	Taiwan	2DS2	Susceptibility	Autoimmune	Arthritis, Rheumatoid
22960345	Mexico	2DS2	Susceptibility	Autoimmune	Arthritis, Rheumatoid
24912006	India	2DS2	Susceptibility	Autoimmune	Arthritis, Rheumatoid
26658904	Iran	2DS4	Susceptibility	Autoimmune	Arthritis, Rheumatoid
26658904	Iran	2DS5	Protection	Autoimmune	Arthritis, Rheumatoid
24912006	India	3DL1	Protection	Autoimmune	Arthritis, Rheumatoid
26658904	Iran	3DL1	Susceptibility	Autoimmune	Arthritis, Rheumatoid
26658904	Iran	3DP1	Susceptibility	Autoimmune	Arthritis, Rheumatoid
26658904	Iran	3DS1	Protection	Autoimmune	Arthritis, Rheumatoid
24912006	India	3DS1	Susceptibility	Autoimmune	Arthritis, Rheumatoid
26334461	Turkey	2DL2	Protection	Autoimmune	Celiac Disease
22180175	Italy	2DL2	Susceptibility	Autoimmune	Celiac Disease
26334461	Turkey	2DP1	Protection	Autoimmune	Celiac Disease
26334461	Turkey	2DS2	Protection	Autoimmune	Celiac Disease
22180175	Italy	2DS2	Susceptibility	Autoimmune	Celiac Disease
26334461	Turkey	2DS4	Protection	Autoimmune	Celiac Disease
26334461	Turkey	2DS5	Susceptibility	Autoimmune	Celiac Disease
22180175	Italy	3DL1	Protection	Autoimmune	Celiac Disease
26334461	Turkey	3DL1	Protection	Autoimmune	Celiac Disease
22180175	Italy	3DL2	Protection	Autoimmune	Celiac Disease
26334461	Turkey	3DS1	Protection	Autoimmune	Celiac Disease
26334461	Turkey	3DS1	Susceptibility	Autoimmune	Celiac Disease
20036705	Brazil	2DL2	Protection	Autoimmune	Colitis, Ulcerative
16929347	United Kingdom	2DL2	Susceptibility	Autoimmune	Colitis, Ulcerative
16929347	United Kingdom	2DS2	Susceptibility	Autoimmune	Colitis, Ulcerative
19789864	United States	2DL2	Protection	Autoimmune	Crohn Disease
26542067	Spain	2DL2	Protection	Autoimmune	Crohn Disease
26542067	Spain	2DL3	Susceptibility	Autoimmune	Crohn Disease
26542067	Spain	2DS2	Protection	Autoimmune	Crohn Disease

15699512	Latvia	2DL1	Protection	Autoimmune	Diabetes Mellitus, Type 1
14514651	Netherlands	2DL1	Protection	Autoimmune	Diabetes Mellitus, Type 1
22069276	China	2DL1	Protection	Autoimmune	Diabetes Mellitus, Type 1
21909837	United Kingdom	2DL1	Protection	Autoimmune	Diabetes Mellitus, Type 1
20580654	Brazil	2DL1	Protection	Autoimmune	Diabetes Mellitus, Type 1
17174747	Finland	2DL1	Susceptibility	Autoimmune	Diabetes Mellitus, Type 1
21909837	United Kingdom	2DL1	Susceptibility	Autoimmune	Diabetes Mellitus, Type 1
17445178	Netherlands	2DL2	Susceptibility	Autoimmune	Diabetes Mellitus, Type 1
20580654	Brazil	2DL2	Susceptibility	Autoimmune	Diabetes Mellitus, Type 1
26031759	India	2DL2	Susceptibility	Autoimmune	Diabetes Mellitus, Type 1
17445178	Netherlands	2DL3	Protection	Autoimmune	Diabetes Mellitus, Type 1
22069276	China	2DL3	Protection	Autoimmune	Diabetes Mellitus, Type 1
21909837	United Kingdom	2DL3	Protection	Autoimmune	Diabetes Mellitus, Type 1
26031759	India	2DL3	Protection	Autoimmune	Diabetes Mellitus, Type 1
17445178	Netherlands	2DL3	Susceptibility	Autoimmune	Diabetes Mellitus, Type 1
22069276	China	2DL3	Susceptibility	Autoimmune	Diabetes Mellitus, Type 1
21909837	United Kingdom	2DL3	Susceptibility	Autoimmune	Diabetes Mellitus, Type 1
26031759	India	2DL3	Susceptibility	Autoimmune	Diabetes Mellitus, Type 1
19046302	Latvia	2DL5	Susceptibility	Autoimmune	Diabetes Mellitus, Type 1
17445178	Netherlands	2DL5	Susceptibility	Autoimmune	Diabetes Mellitus, Type 1
21909837	United Kingdom	2DS1	Protection	Autoimmune	Diabetes Mellitus, Type 1
19046302	Latvia	2DS1	Susceptibility	Autoimmune	Diabetes Mellitus, Type 1
17445178	Netherlands	2DS1	Susceptibility	Autoimmune	Diabetes Mellitus, Type 1
21909837	United Kingdom	2DS1	Susceptibility	Autoimmune	Diabetes Mellitus, Type 1
19046302	Latvia	2DS2	Susceptibility	Autoimmune	Diabetes Mellitus, Type 1
17445178	Netherlands	2DS2	Susceptibility	Autoimmune	Diabetes Mellitus, Type 1
26031759	India	2DS2	Susceptibility	Autoimmune	Diabetes Mellitus, Type 1
19046302	Latvia	2DS3	Susceptibility	Autoimmune	Diabetes Mellitus, Type 1
17445178	Netherlands	2DS3	Susceptibility	Autoimmune	Diabetes Mellitus, Type 1
17174747	Finland	2DS4	Susceptibility	Autoimmune	Diabetes Mellitus, Type 1
26031759	India	2DS4	Susceptibility	Autoimmune	Diabetes Mellitus, Type 1
14514651	Netherlands	3DL1	Protection	Autoimmune	Diabetes Mellitus, Type 1
15699512	Latvia	3DS1	Susceptibility	Autoimmune	Diabetes Mellitus, Type 1
19046302	Latvia	3DS1	Susceptibility	Autoimmune	Diabetes Mellitus, Type 1
17445178	Netherlands	3DS1	Susceptibility	Autoimmune	Diabetes Mellitus, Type 1
20580654	Brazil	3DS1	Susceptibility	Autoimmune	Diabetes Mellitus, Type 1
19309280	Lebanon	2DP1	Protection	Autoimmune	Familial Mediterranean Fever
26574972	Turkey	2DS2	Susceptibility	Autoimmune	Familial Mediterranean Fever
24367901	Australia	2DL1	Susceptibility	Autoimmune	Guillain Barre Syndrome
27797112	Spain	2DL5	Susceptibility	Autoimmune	Inflammatory Bowel Diseases
27797112	Spain	2DS1	Susceptibility	Autoimmune	Inflammatory Bowel Diseases
27797112	Spain	2DS5	Susceptibility	Autoimmune	Inflammatory Bowel Diseases
27797112	Spain	3DS1	Susceptibility	Autoimmune	Inflammatory Bowel Diseases
19120281		2DL1	Susceptibility	Autoimmune	Latent Autoimmune Diabetes in Adults
19120281		2DL2	Protection	Autoimmune	Latent Autoimmune Diabetes in Adults
19120281		2DL2	Susceptibility	Autoimmune	Latent Autoimmune Diabetes in Adults

19120281		2DL5	Susceptibility	Autoimmune	Latent Autoimmune Diabetes in Adults
19120281		2DS1	Protection	Autoimmune	Latent Autoimmune Diabetes in Adults
19120281		2DS2	Susceptibility	Autoimmune	Latent Autoimmune Diabetes in Adults
19120281		2DS3	Protection	Autoimmune	Latent Autoimmune Diabetes in Adults
19120281		2DS4	Susceptibility	Autoimmune	Latent Autoimmune Diabetes in Adults
19120281		2DS5	Protection	Autoimmune	Latent Autoimmune Diabetes in Adults
19120281		3DL1	Susceptibility	Autoimmune	Latent Autoimmune Diabetes in Adults
19926642	China	2DL2	Susceptibility	Autoimmune	Lupus Erythematosus, Systemic
20371502	Japan	2DL5	Protection	Autoimmune	Lupus Erythematosus, Systemic
17445179	Canada	2DS1	Susceptibility	Autoimmune	Lupus Erythematosus, Systemic
20371502	Japan	2DS1	Susceptibility	Autoimmune	Lupus Erythematosus, Systemic
19926642	China	2DS1	Susceptibility	Autoimmune	Lupus Erythematosus, Systemic
25581336	China	2DS1	Susceptibility	Autoimmune	Lupus Erythematosus, Systemic
26989167	Iran	3DP1	Susceptibility	Autoimmune	Lupus Erythematosus, Systemic
16508981	Japan	2DS3	Protection	Autoimmune	Microscopic Polyangiitis
19630074	Norway	2DL1	Protection	Autoimmune	Multiple Sclerosis
22185807	Germany	2DL2	Susceptibility	Autoimmune	Multiple Sclerosis
24735502	Tunisia	2DL2	Susceptibility	Autoimmune	Multiple Sclerosis
22185807	Germany	2DL3	Protection	Autoimmune	Multiple Sclerosis
21665278	Spain	2DL5	Susceptibility	Autoimmune	Multiple Sclerosis
19630074	Norway	2DS1	Protection	Autoimmune	Multiple Sclerosis
24529855	Portugal	2DS1	Protection	Autoimmune	Multiple Sclerosis
22185807	Germany	2DS2	Susceptibility	Autoimmune	Multiple Sclerosis
19630074	Norway	2DS4	Susceptibility	Autoimmune	Multiple Sclerosis
19630074	Norway	3DL1	Susceptibility	Autoimmune	Multiple Sclerosis
21665278	Spain	3DS1	Susceptibility	Autoimmune	Multiple Sclerosis
22768326	Brazil	2DL5	Protection	Autoimmune	Pemphigus Foliaceus
22768326	Brazil	2DS1	Protection	Autoimmune	Pemphigus Foliaceus
22768326	Brazil	2DS3	Protection	Autoimmune	Pemphigus Foliaceus
25867094	Brazil	3DL2	Susceptibility	Autoimmune	Pemphigus Foliaceus
22768326	Brazil	3DS1	Protection	Autoimmune	Pemphigus Foliaceus
26198918	Australia	3DL1	Susceptibility	Autoimmune	Polyradiculoneuropathy, Chronic Inflammatory Demyelinating
15140215	Japan	2DL5	Susceptibility	Autoimmune	Psoriasis
18643961	Brazil	2DS1	Susceptibility	Autoimmune	Psoriasis
16829306	Poland	2DS1	Susceptibility	Autoimmune	Psoriasis
16185272	Sweden	2DS1	Susceptibility	Autoimmune	Psoriasis
15140215	Japan	2DS1	Susceptibility	Autoimmune	Psoriasis
22024796	Australia	2DL2	Susceptibility	Autoimmune	Purpura, Thrombocytopenic, Idiopathic
24571473	Australia	2DL2	Susceptibility	Autoimmune	Purpura, Thrombocytopenic, Idiopathic
24571473	Australia	2DL5	Protection	Autoimmune	Purpura, Thrombocytopenic, Idiopathic

22024796	Australia	2DS2	Susceptibility	Autoimmune	Purpura, Thrombocytopenic, Idiopathic
24571473	Australia	2DS2	Susceptibility	Autoimmune	Purpura, Thrombocytopenic, Idiopathic
24571473	Australia	2DS3	Susceptibility	Autoimmune	Purpura, Thrombocytopenic, Idiopathic
24571473	Australia	2DS5	Protection	Autoimmune	Purpura, Thrombocytopenic, Idiopathic
20082621	Brazil	2DL2	Protection	Autoimmune	Scleroderma, Systemic
17445179	Canada	2DL2	Susceptibility	Autoimmune	Scleroderma, Systemic
17445179	Canada	2DS1	Protection	Autoimmune	Scleroderma, Systemic
17445179	Canada	2DS1	Susceptibility	Autoimmune	Scleroderma, Systemic
26996109	Iran	2DL1	Protection	Autoimmune	Spondylitis, Ankylosing
20652381	China	2DL1	Susceptibility	Autoimmune	Spondylitis, Ankylosing
26996109	Iran	2DL1	Susceptibility	Autoimmune	Spondylitis, Ankylosing
26996109	Iran	2DL2	Protection	Autoimmune	Spondylitis, Ankylosing
26996109	Iran	2DL3	Susceptibility	Autoimmune	Spondylitis, Ankylosing
26996109	Iran	2DL5	Protection	Autoimmune	Spondylitis, Ankylosing
21797986	Iran	2DL5	Susceptibility	Autoimmune	Spondylitis, Ankylosing
20652381	China	2DL5	Susceptibility	Autoimmune	Spondylitis, Ankylosing
26996109	Iran	2DS1	Protection	Autoimmune	Spondylitis, Ankylosing
21797986	Iran	2DS1	Susceptibility	Autoimmune	Spondylitis, Ankylosing
25491925	Spain	2DS1	Susceptibility	Autoimmune	Spondylitis, Ankylosing
26996109	Iran	2DS1	Susceptibility	Autoimmune	Spondylitis, Ankylosing
26996109	Iran	2DS2	Protection	Autoimmune	Spondylitis, Ankylosing
21797986	Iran	3DL1	Protection	Autoimmune	Spondylitis, Ankylosing
20574122	Iran	3DL1	Protection	Autoimmune	Spondylitis, Ankylosing
22744805	China	3DS1	Susceptibility	Autoimmune	Spondylitis, Ankylosing
21797986	Iran	3DS1	Susceptibility	Autoimmune	Spondylitis, Ankylosing
25491925	Spain	3DS1	Susceptibility	Autoimmune	Spondylitis, Ankylosing
27490240	Japan	2DL2	Protection	Autoimmune	Uveomeningoencephalitic Syndrome
19897003	Japan	2DS1	Susceptibility	Autoimmune	Uveomeningoencephalitic Syndrome
19897003	Japan	2DS2	Protection	Autoimmune	Uveomeningoencephalitic Syndrome
27490240	Japan	2DS2	Protection	Autoimmune	Uveomeningoencephalitic Syndrome
27490240	Japan	2DS3	Protection	Autoimmune	Uveomeningoencephalitic Syndrome
22219647	Saudi Arabia	2DS3	Susceptibility	Autoimmune	Uveomeningoencephalitic Syndrome
19897003	Japan	2DS5	Susceptibility	Autoimmune	Uveomeningoencephalitic Syndrome
19897003	Japan	3DL1	Protection	Autoimmune	Uveomeningoencephalitic Syndrome
27490240	Japan	3DL1	Protection	Autoimmune	Uveomeningoencephalitic Syndrome
21479698	Turkey	2DL1	Protection	Cancer	Breast Neoplasms
27631728	Saudi Arabia	2DL1	Protection	Cancer	Breast Neoplasms
27631728	Saudi Arabia	2DL1	Susceptibility	Cancer	Breast Neoplasms
27631728	Saudi Arabia	2DL2	Protection	Cancer	Breast Neoplasms
23792055	Brazil	2DL2	Susceptibility	Cancer	Breast Neoplasms
27631728	Saudi Arabia	2DL3	Protection	Cancer	Breast Neoplasms
27631728	Saudi Arabia	2DL5	Protection	Cancer	Breast Neoplasms
21479698	Turkey	2DS1	Susceptibility	Cancer	Breast Neoplasms
27631728	Saudi Arabia	2DS2	Protection	Cancer	Breast Neoplasms
27631728	Saudi Arabia	2DS3	Protection	Cancer	Breast Neoplasms

25700262	Italy	2DS4	Susceptibility	Cancer	Carcinoma, Hepatocellular
25700262	Italy	3DL1	Susceptibility	Cancer	Carcinoma, Hepatocellular
24011088	Australia	2DL2	Protection	Cancer	Cervical Intraepithelial Neoplasia
24011088	Australia	2DS2	Protection	Cancer	Cervical Intraepithelial Neoplasia
26181663	Turkey	2DL1	Protection	Cancer	Colorectal Neoplasms
28088355	Brazil	2DL1	Protection	Cancer	Colorectal Neoplasms
27519478	Italy	2DL2	Protection	Cancer	Colorectal Neoplasms
26181663	Turkey	2DL2	Susceptibility	Cancer	Colorectal Neoplasms
26383988	Saudi Arabia	2DL3	Susceptibility	Cancer	Colorectal Neoplasms
26383988	Saudi Arabia	2DS1	Susceptibility	Cancer	Colorectal Neoplasms
26383988	Saudi Arabia	2DS2	Susceptibility	Cancer	Colorectal Neoplasms
26383988	Saudi Arabia	2DS3	Susceptibility	Cancer	Colorectal Neoplasms
28088355	Brazil	2DS4	Protection	Cancer	Colorectal Neoplasms
24998207	South Korea	2DS5	Susceptibility	Cancer	Colorectal Neoplasms
26383988	Saudi Arabia	2DS5	Susceptibility	Cancer	Colorectal Neoplasms
24755352	Thailand	3DL1	Protection	Cancer	Diffuse Large B-cell Lymphoma
24755350	Poland	2DL1	Protection	Cancer	Epithelial Ovarian Cancer
24755350	Poland	2DL3	Susceptibility	Cancer	Epithelial Ovarian Cancer
26495028	Mexico	2DL2	Susceptibility	Cancer	Haematological Malignancies
26983546	Italy	3DL1	Susceptibility	Cancer	Hodgkin Disease
26983546	Italy	3DS1	Susceptibility	Cancer	Hodgkin Disease
21726204	Poland	3DL1	Protection	Cancer	Leukemia, Lymphocytic, Chronic, B-Cell
21726204	Poland	3DS1	Protection	Cancer	Leukemia, Lymphocytic, Chronic, B-Cell
26472014	India	2DL1	Protection	Cancer	Lymphoblastic Leukemia, Acute
26472014	India	2DL2	Protection	Cancer	Lymphoblastic Leukemia, Acute
26472014	India	2DL3	Protection	Cancer	Lymphoblastic Leukemia, Acute
26472014	India	2DL5	Protection	Cancer	Lymphoblastic Leukemia, Acute
25281696	United States	2DL5	Susceptibility	Cancer	Lymphoblastic Leukemia, Acute
26472014	India	2DS1	Susceptibility	Cancer	Lymphoblastic Leukemia, Acute
25281696	United States	2DS1	Susceptibility	Cancer	Lymphoblastic Leukemia, Acute
26472014	India	2DS2	Susceptibility	Cancer	Lymphoblastic Leukemia, Acute
26472014	India	2DS3	Susceptibility	Cancer	Lymphoblastic Leukemia, Acute
25281696	United States	2DS3	Susceptibility	Cancer	Lymphoblastic Leukemia, Acute
26472014	India	2DS4	Susceptibility	Cancer	Lymphoblastic Leukemia, Acute
26472014	India	2DS5	Susceptibility	Cancer	Lymphoblastic Leukemia, Acute
26472014	India	3DL1	Protection	Cancer	Lymphoblastic Leukemia, Acute
24518758	United States	3DL1	Susceptibility	Cancer	Lymphoblastic Leukemia, Acute
26472014	India	3DS1	Susceptibility	Cancer	Lymphoblastic Leukemia, Acute
24518758	United States	3DS1	Susceptibility	Cancer	Lymphoblastic Leukemia, Acute
25699652	Turkey	2DS1	Susceptibility	Cancer	Lymphoma, Non-Hodgkin

25699652	Turkey	3DL1	Susceptibility	Cancer	Lymphoma, Non-Hodgkin
25699652	Turkey	3DS1	Protection	Cancer	Lymphoma, Non-Hodgkin
27141379	Spain	2DL1	Protection	Cancer	Multiple Myeloma
27141379	Spain	2DL3	Protection	Cancer	Multiple Myeloma
27141379	Spain	2DS4	Protection	Cancer	Multiple Myeloma
27141379	Spain	3DL1	Protection	Cancer	Multiple Myeloma
26202659	Ireland	2DL2	Susceptibility	Cancer	Neuroblastoma
26202659	Ireland	2DS2	Susceptibility	Cancer	Neuroblastoma
24818561	India	2DL3	Protection	Cancer	Oral Squamous Cell Carcinoma
26268853	Italy	3DS1	Susceptibility	Cancer	Sarcoma, Kaposi
25978047	Brazil	2DL2	Protection	Infectious	Chagas Disease
25978047	Brazil	3DL2	Protection	Infectious	Chagas Disease
24737799	Italy	2DS2	Protection	Infectious	Cytomegalovirus Infections
24498996	Brazil	2DL5	Susceptibility	Infectious	Dengue
26385514	India	3DL1	Susceptibility	Infectious	Dengue
25966757	China	2DL2	Susceptibility	Infectious	Epstein-Barr Virus Infections
25966757	China	2DL5	Susceptibility	Infectious	Epstein-Barr Virus Infections
25966757	China	2DS2	Susceptibility	Infectious	Epstein-Barr Virus Infections
25966757	China	2DS4	Susceptibility	Infectious	Epstein-Barr Virus Infections
25966757	China	3DS1	Protection	Infectious	Epstein-Barr Virus Infections
20878400	Gabon	2DS1	Susceptibility	Infectious	Hemorrhagic Fever, Ebola
20878400	Gabon	2DS3	Protection	Infectious	Hemorrhagic Fever, Ebola
20878400	Gabon	2DS3	Susceptibility	Infectious	Hemorrhagic Fever, Ebola
20878400	Gabon	2DS4	Protection	Infectious	Hemorrhagic Fever, Ebola
24407110	Turkey	2DL3	Protection	Infectious	Hepatitis B
18074414	China	2DL5	Protection	Infectious	Hepatitis B
18074414	China	2DL5	Susceptibility	Infectious	Hepatitis B
18074414	China	2DS1	Protection	Infectious	Hepatitis B
18074414	China	2DS2	Susceptibility	Infectious	Hepatitis B
18074414	China	2DS3	Susceptibility	Infectious	Hepatitis B
27019428	India	2DS5	Protection	Infectious	Hepatitis B
18074414	China	2DS5	Susceptibility	Infectious	Hepatitis B
18074414	China	3DS1	Protection	Infectious	Hepatitis B
24407110	Turkey	3DS1	Protection	Infectious	Hepatitis B
26945896	Netherlands	2DL2	Protection	Infectious	Hepatitis B, Chronic
26945896	Netherlands	2DL3	Susceptibility	Infectious	Hepatitis B, Chronic
20643584	China	2DL1	Susceptibility	Infectious	Hepatitis C
16571411	Spain	2DL2	Protection	Infectious	Hepatitis C
17445180	Argentina	2DL2	Protection	Infectious	Hepatitis C
21931540	Australia	2DL2	Protection	Infectious	Hepatitis C
17445180	Argentina	2DL2	Susceptibility	Infectious	Hepatitis C
19846535	Spain	2DL2	Susceptibility	Infectious	Hepatitis C
23569404	Brazil	2DL2	Susceptibility	Infectious	Hepatitis C
20077564	United Kingdom	2DL3	Protection	Infectious	Hepatitis C
19846535	Spain	2DL3	Protection	Infectious	Hepatitis C
21931540	Australia	2DL3	Protection	Infectious	Hepatitis C
20643584	China	2DL3	Protection	Infectious	Hepatitis C
18289678	United States	2DL3	Protection	Infectious	Hepatitis C
16571411	Spain	2DL3	Susceptibility	Infectious	Hepatitis C
21931540	Australia	2DL3	Susceptibility	Infectious	Hepatitis C
18289678	United States	2DL3	Susceptibility	Infectious	Hepatitis C
20456039	Brazil	2DL5	Susceptibility	Infectious	Hepatitis C
21931540	Australia	2DS1	Protection	Infectious	Hepatitis C
17445180	Argentina	2DS2	Susceptibility	Infectious	Hepatitis C

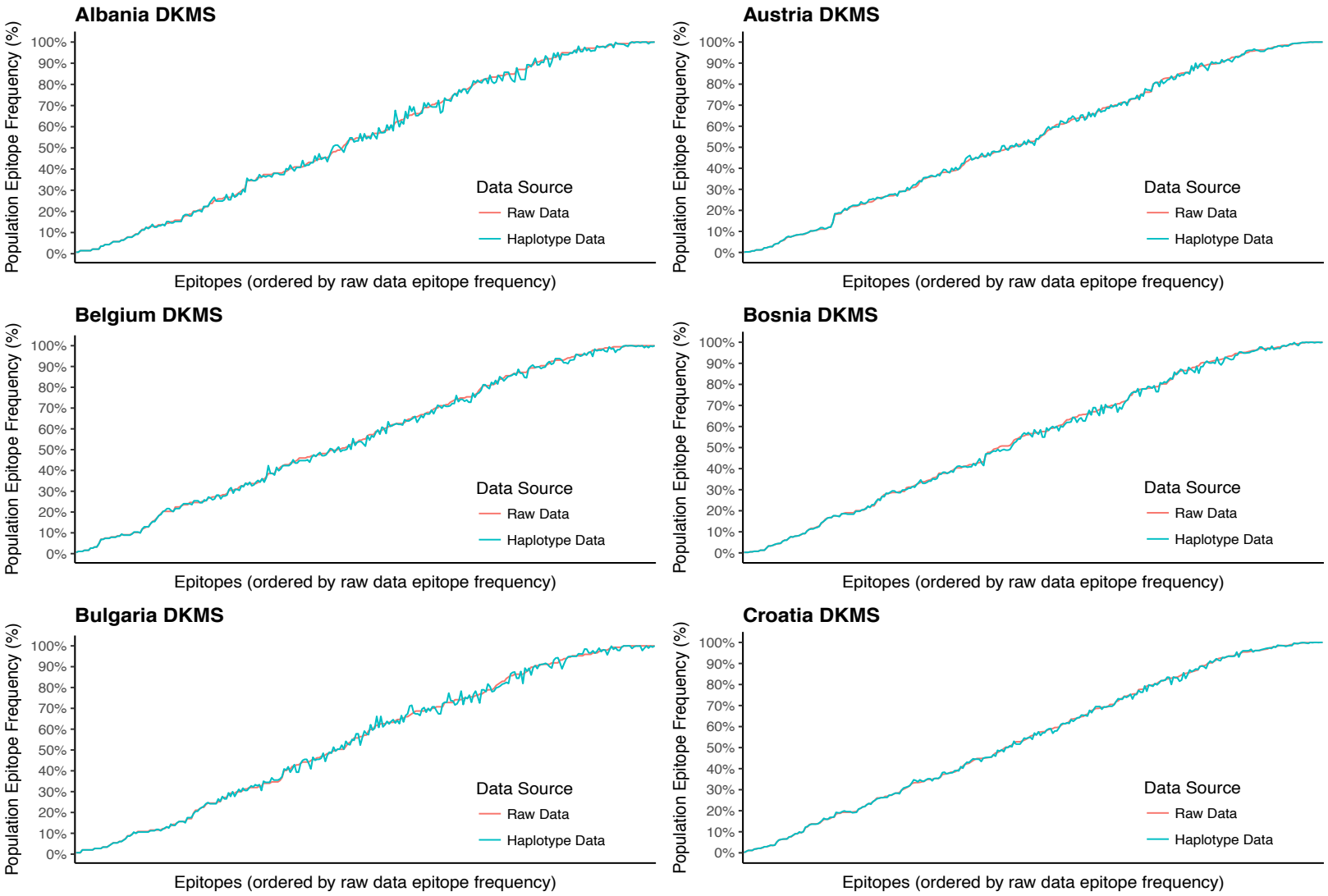
23569404	Brazil	2DS2	Susceptibility	Infectious	Hepatitis C
17445180	Argentina	2DS3	Susceptibility	Infectious	Hepatitis C
21402922	Ireland	2DS3	Susceptibility	Infectious	Hepatitis C
23569404	Brazil	2DS3	Susceptibility	Infectious	Hepatitis C
17445180	Argentina	2DS4	Protection	Infectious	Hepatitis C
19552960	Puerto Rico	2DS4	Protection	Infectious	Hepatitis C
17445180	Argentina	2DS5	Susceptibility	Infectious	Hepatitis C
20456039	Brazil	2DS5	Susceptibility	Infectious	Hepatitis C
17445180	Argentina	3DL1	Protection	Infectious	Hepatitis C
15942906	Spain	3DS1	Protection	Infectious	Hepatitis C
25636579	Poland	2DS2	Protection	Infectious	Hepatitis C, Chronic
17559579	Spain	2DL2	Susceptibility	Infectious	Herpes Simplex
17559579	Spain	2DS2	Susceptibility	Infectious	Herpes Simplex
27930093	Zimbabwe	2DL2	Protection	Infectious	HIV Infections
27148256	China	2DL2	Protection	Infectious	HIV Infections
22073315	Italy	2DL3	Protection	Infectious	HIV Infections
26888639	Poland	2DL3	Protection	Infectious	HIV Infections
26255774	India	2DL3	Protection	Infectious	HIV Infections
17082569	Cote d Ivoire	2DL3	Susceptibility	Infectious	HIV Infections
27148256	China	2DL5	Protection	Infectious	HIV Infections
26888639	Poland	2DL5	Susceptibility	Infectious	HIV Infections
22073315	Italy	2DS1	Susceptibility	Infectious	HIV Infections
27148256	China	2DS2	Protection	Infectious	HIV Infections
27148256	China	2DS4	Protection	Infectious	HIV Infections
26255774	India	2DS5	Protection	Infectious	HIV Infections
27148256	China	2DS5	Protection	Infectious	HIV Infections
15784466	Zambia	3DL1	Protection	Infectious	HIV Infections
17082569	Cote d Ivoire	3DL1	Protection	Infectious	HIV Infections
26033692	Mexico	3DL1	Protection	Infectious	HIV Infections
27148256	China	3DL1	Protection	Infectious	HIV Infections
18317000	Canada	3DS1	Protection	Infectious	HIV Infections
24877146	India	3DS1	Protection	Infectious	HIV Infections
26033692	Mexico	3DS1	Protection	Infectious	HIV Infections
18778326	Brazil	2DL1	Protection	Infectious	Leprosy
25117794	Brazil	2DL1	Protection	Infectious	Leprosy
18778326	Brazil	2DL3	Protection	Infectious	Leprosy
25117794	Brazil	2DS2	Susceptibility	Infectious	Leprosy
18778326	Brazil	2DS3	Susceptibility	Infectious	Leprosy
18778326	Brazil	3DL2	Susceptibility	Infectious	Leprosy
22715396	Kenya	2DL1	Susceptibility	Infectious	Malaria
22715396	Kenya	2DL2	Protection	Infectious	Malaria
22715396	Kenya	2DL2	Susceptibility	Infectious	Malaria
22715396	Kenya	2DL3	Protection	Infectious	Malaria
22412373	Thailand	2DL3	Susceptibility	Infectious	Malaria
22715396	Kenya	2DL3	Susceptibility	Infectious	Malaria
24929143	Nigeria	2DL5	Protection	Infectious	Malaria
22412373	Thailand	2DS1	Protection	Infectious	Malaria
22412373	Thailand	2DS1	Susceptibility	Infectious	Malaria
24929143	Nigeria	2DS3	Protection	Infectious	Malaria
24929143	Nigeria	2DS3	Protection	Infectious	Malaria
19859704	Solomon Islands	2DS4	Susceptibility	Infectious	Malaria
24929143	Nigeria	2DS5	Protection	Infectious	Malaria
24929143	Nigeria	2DS5	Protection	Infectious	Malaria
19859704	Solomon Islands	3DL1	Susceptibility	Infectious	Malaria
21889618	India	3DL1	Susceptibility	Infectious	Malaria
25188020	Italy	2DL3	Protection	Infectious	Papillomavirus Infections
22958291	china	2DL3	Protection	Infectious	Syphilis
22958291	china	2DS3	Susceptibility	Infectious	Syphilis
22128817	China	2DS4	Susceptibility	Infectious	Syphilis
22958291	china	2DS5	Protection	Infectious	Syphilis

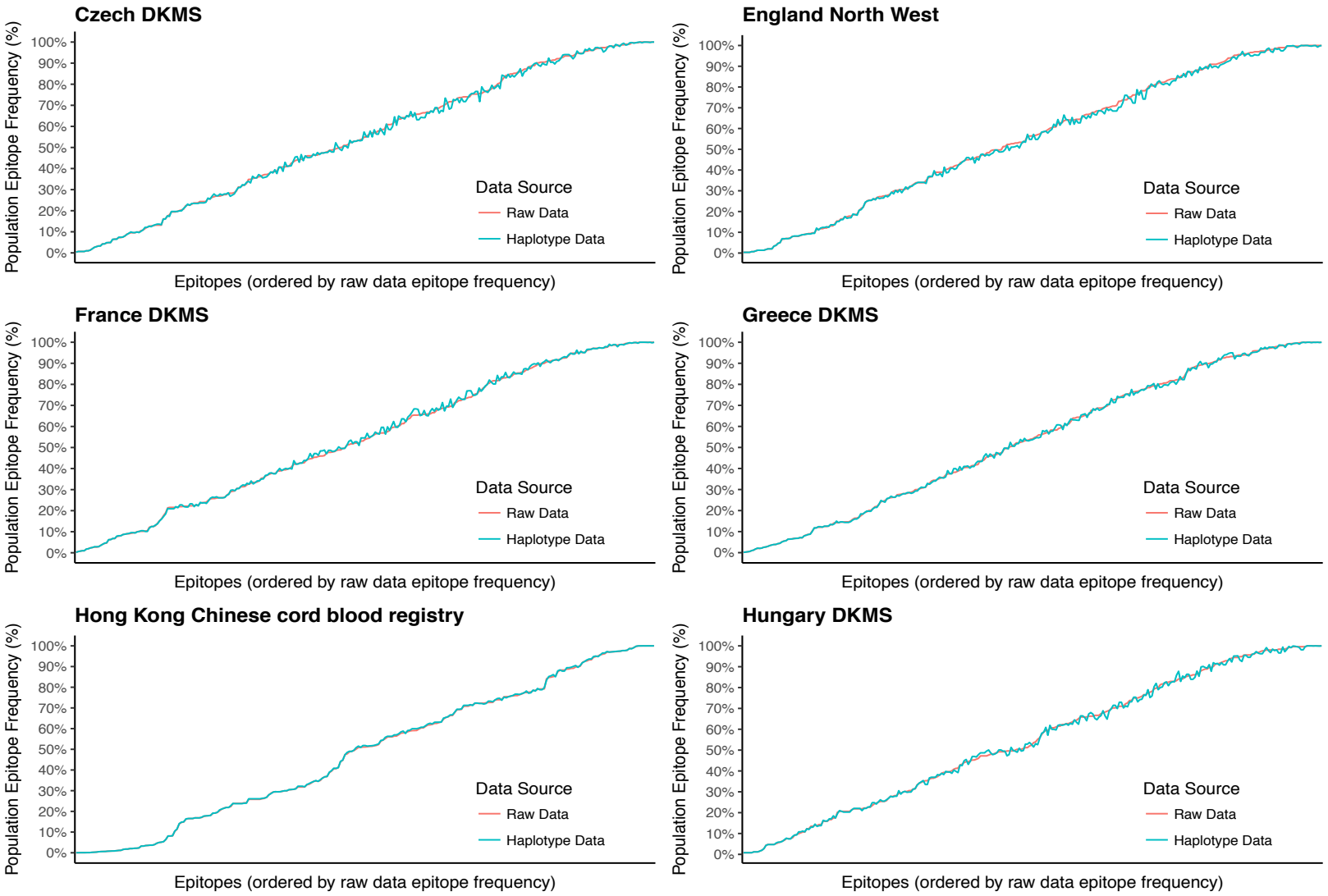
22958291	china	3DS1	Susceptibility	Infectious	Syphilis
26077983	Canada	2DL2	Protection	Infectious	Tuberculosis
23073291	India	2DL3	Susceptibility	Infectious	Tuberculosis
17092251	Mexico	2DS1	Susceptibility	Infectious	Tuberculosis
23073291	India	2DS1	Susceptibility	Infectious	Tuberculosis
26077983	Canada	2DS2	Protection	Infectious	Tuberculosis
23073291	India	2DS5	Susceptibility	Infectious	Tuberculosis
23073291	India	3DL1	Susceptibility	Infectious	Tuberculosis
22426166	Iran	3DS1	Protection	Infectious	Tuberculosis
22862677	China	2DL2	Protection	Infectious	Tuberculosis, Pulmonary
22862677	China	2DL2	Susceptibility	Infectious	Tuberculosis, Pulmonary
22862677	China	2DL3	Protection	Infectious	Tuberculosis, Pulmonary
22862677	China	2DL3	Susceptibility	Infectious	Tuberculosis, Pulmonary
22862677	China	2DS1	Susceptibility	Infectious	Tuberculosis, Pulmonary
22653583	China	2DS1	Susceptibility	Infectious	Tuberculosis, Pulmonary
22862677	China	2DS3	Susceptibility	Infectious	Tuberculosis, Pulmonary
22653583	China	2DS3	Susceptibility	Infectious	Tuberculosis, Pulmonary
22862677	China	3DS1	Susceptibility	Infectious	Tuberculosis, Pulmonary
22653583	China	3DS1	Susceptibility	Infectious	Tuberculosis, Pulmonary
22884899	United States	2DL5	Susceptibility	Neurological	Autism
22884899	United States	2DS1	Susceptibility	Neurological	Autism
24120931	Italy	2DS2	Susceptibility	Neurological	Autism
22884899	United States	2DS4	Susceptibility	Neurological	Autism
22884899	United States	2DS5	Susceptibility	Neurological	Autism
22884899	United States	3DL1	Protection	Neurological	Autism
24120931	Italy	3DL1	Protection	Neurological	Autism
22884899	United States	3DL1	Susceptibility	Neurological	Autism
22884899	United States	3DS1	Susceptibility	Neurological	Autism
21159685	India	2DL1	Protection	Pregnancy	Abortion, Habitual
19279038	India	2DL1	Protection	Pregnancy	Abortion, Habitual
17617375	China	2DL5	Susceptibility	Pregnancy	Abortion, Habitual
19279038	India	2DP1	Protection	Pregnancy	Abortion, Habitual
17617375	China	2DS1	Susceptibility	Pregnancy	Abortion, Habitual
17617375	China	2DS2	Susceptibility	Pregnancy	Abortion, Habitual
21159685	India	2DS2	Susceptibility	Pregnancy	Abortion, Habitual
19279038	India	2DS2	Susceptibility	Pregnancy	Abortion, Habitual
19279038	India	2DS3	Susceptibility	Pregnancy	Abortion, Habitual
21159685	India	2DS4	Protection	Pregnancy	Abortion, Habitual
19279038	India	2DS4	Protection	Pregnancy	Abortion, Habitual
21159685	India	2DS5	Susceptibility	Pregnancy	Abortion, Habitual
19279038	India	3DL1	Protection	Pregnancy	Abortion, Habitual
19279038	India	3DS1	Susceptibility	Pregnancy	Abortion, Habitual
28069185	United States	2DS1	Susceptibility	Pregnancy	Abortion, Spontaneous
28069185	United States	2DS5	Susceptibility	Pregnancy	Abortion, Spontaneous
28069185	United States	3DS1	Susceptibility	Pregnancy	Abortion, Spontaneous
25561558	Uganda	2DL2	Protection	Pregnancy	Pre-Eclampsia
26823774	China	2DL4	Protection	Pregnancy	Pre-Eclampsia
25561558	Uganda	2DL5	Protection	Pregnancy	Pre-Eclampsia
24911933	China	2DS1	Protection	Pregnancy	Pre-Eclampsia
25561558	Uganda	2DS5	Protection	Pregnancy	Pre-Eclampsia
21561401	United States	2DL1	Protection	Undetermined	Bone Marrow failure syndromes
21561401	United States	2DL3	Protection	Undetermined	Bone Marrow failure syndromes
21561401	United States	3DL1	Protection	Undetermined	Bone Marrow failure syndromes
27665490	United States	2DL3	Susceptibility	Undetermined	Common Variable Immunodeficiency
27665490	United States	2DL5	Susceptibility	Undetermined	Common Variable Immunodeficiency
27665490	United States	2DS1	Susceptibility	Undetermined	Common Variable Immunodeficiency

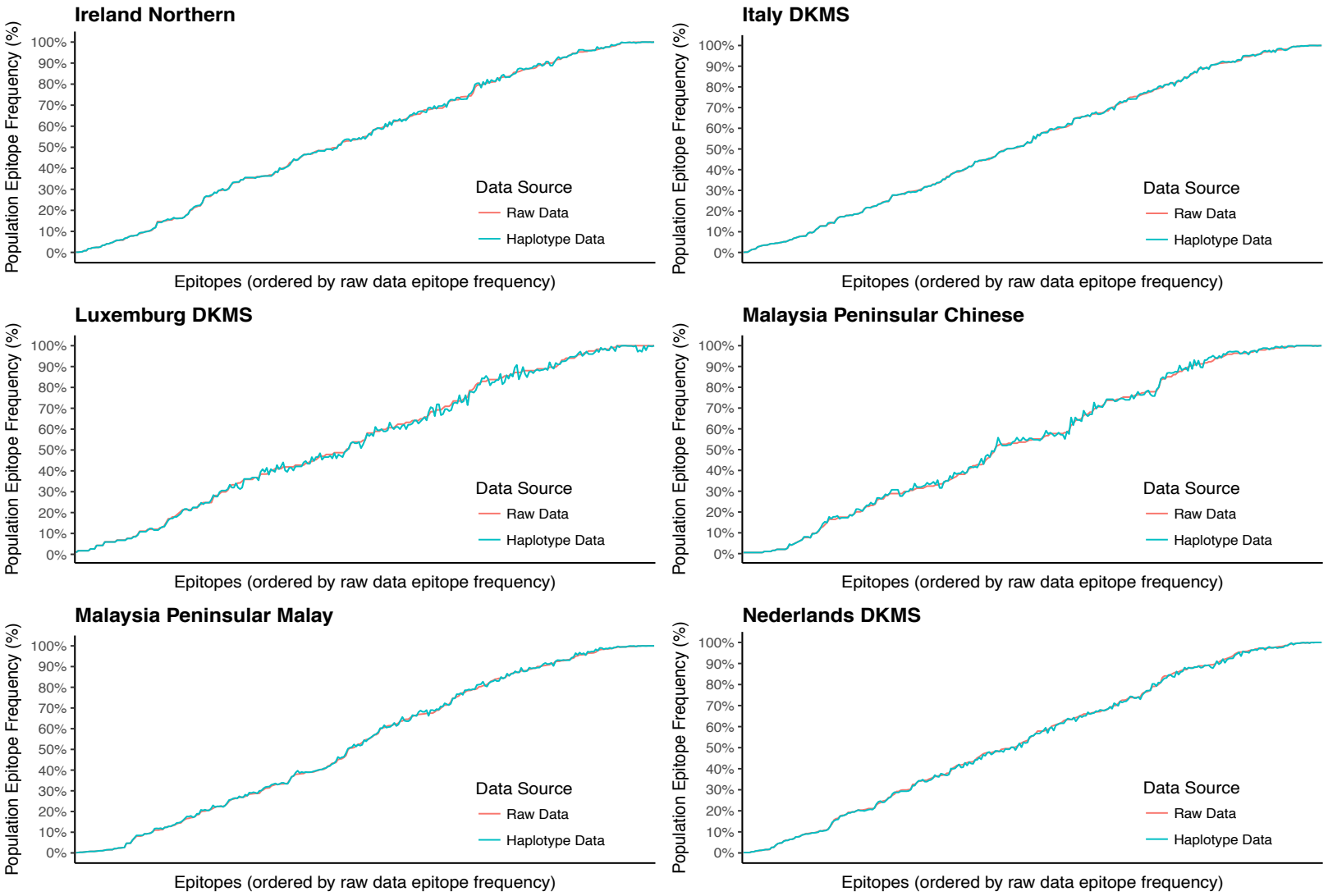
27665490	United States	3DL1	Protection	Undetermined	Common Variable Immunodeficiency
27665490	United States	3DS1	Susceptibility	Undetermined	Common Variable Immunodeficiency
26679162		2DL2	Protection	Undetermined	Cryptorchidism
26679162		2DS2	Protection	Undetermined	Cryptorchidism
23831511	Poland	2DS1	Protection	Undetermined	Dermatitis, Atopic
24122895	Italy	2DL2	Susceptibility	Undetermined	Diabetes Mellitus, Type 2
24122895	Italy	2DL3	Protection	Undetermined	Diabetes Mellitus, Type 2
24122895	Italy	2DS2	Susceptibility	Undetermined	Diabetes Mellitus, Type 2
22509813	China	2DS2	Susceptibility	Undetermined	Dry Eye Syndrome
25724317	Poland	2DS5	Protection	Undetermined	Endometriosis
21468604	Italy	3DL1	Susceptibility	Undetermined	Fatigue Syndrome, Chronic
21468604	Italy	3DS1	Susceptibility	Undetermined	Fatigue Syndrome, Chronic
22803950	Italy	3DL1	Protection	Undetermined	Hemoglobinuria, Paroxysmal
23777934	India	2DL1	Protection	Undetermined	Renal Disease
23777934	India	2DL2	Protection	Undetermined	Renal Disease
23777934	India	2DP1	Protection	Undetermined	Renal Disease
23777934	India	2DS1	Susceptibility	Undetermined	Renal Disease
23777934	India	2DS2	Susceptibility	Undetermined	Renal Disease
23777934	India	2DS3	Susceptibility	Undetermined	Renal Disease
23777934	India	2DS5	Susceptibility	Undetermined	Renal Disease
23777934	India	3DL1	Protection	Undetermined	Renal Disease
23777934	India	3DS1	Susceptibility	Undetermined	Renal Disease
19850842	United States	2DS5	Protection	Undetermined	Uveitis, Anterior
19850842	United States	3DL1	Protection	Undetermined	Uveitis, Anterior
19850842	United States	3DL1	Susceptibility	Undetermined	Uveitis, Anterior

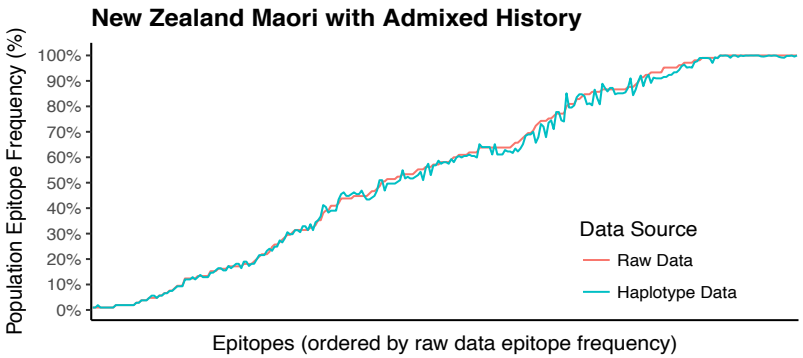
Appendix C

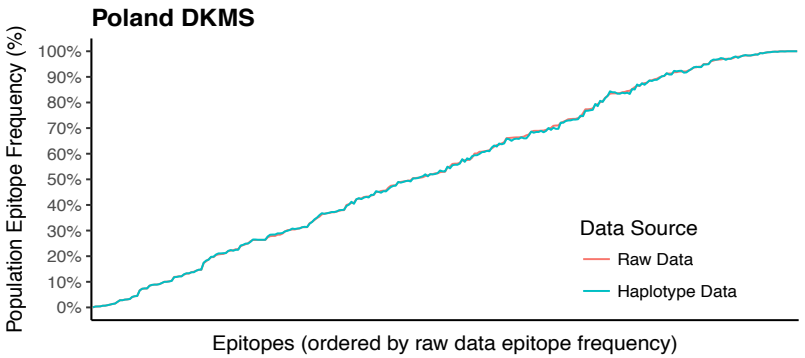
The following figures complement Figure 3.9 by showing more detailed comparisons between HLA epitope frequencies generated from HLA raw data and HLA haplotype frequency data in line plots for each population included in the analysis.

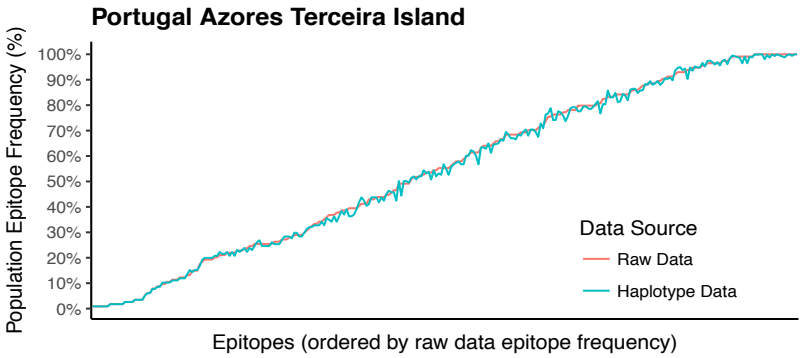

Albania DKMS
Austria DKMS
Belgium DKMS
Bosnia DKMS
Bulgaria DKMS
Croatia DKMS
Population Epitope Frequency (%)
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Data Source
Raw Data
Haplotype Data
Epitopes (ordered by raw data epitope frequency)

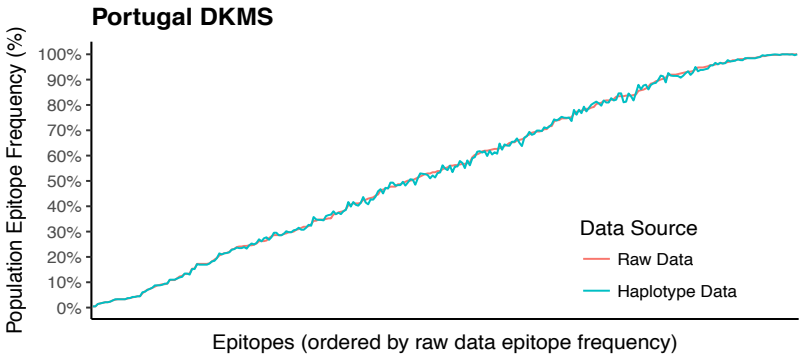

Czech DKMS
England North West
France DKMS
Greece DKMS
Hong Kong Chinese cord blood registry
Hungary DKMS
Population Epitope Frequency (%)
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Data Source
Raw Data
Haplotype Data
Epitopes (ordered by raw data epitope frequency)

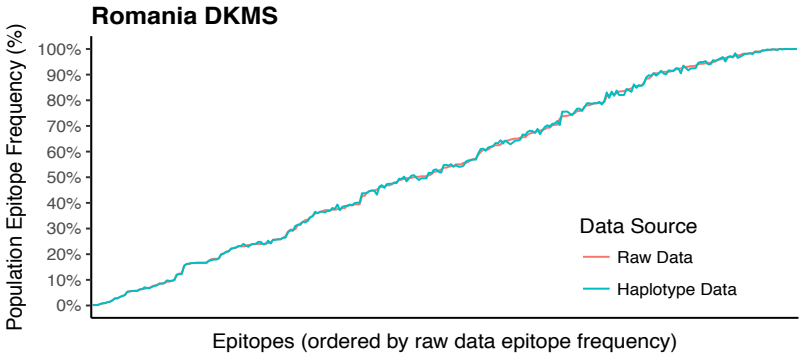

Ireland Northern
Italy DKMS
Luxemburg DKMS
Malaysia Peninsular Chinese
Malaysia Peninsular Malay
Nederlands DKMS
Population Epitope Frequency (%)
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Data Source
Raw Data
Haplotype Data
Epitopes (ordered by raw data epitope frequency)

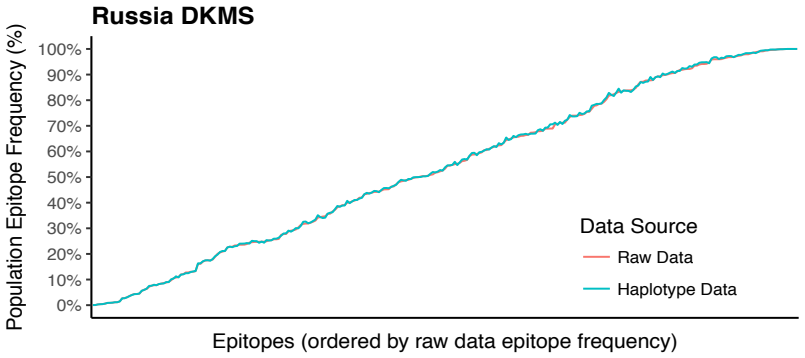

New Zealand Maori with Admixed History
Population Epitope Frequency (%)
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Data Source
Raw Data
Haplotype Data
Epitopes (ordered by raw data epitope frequency)

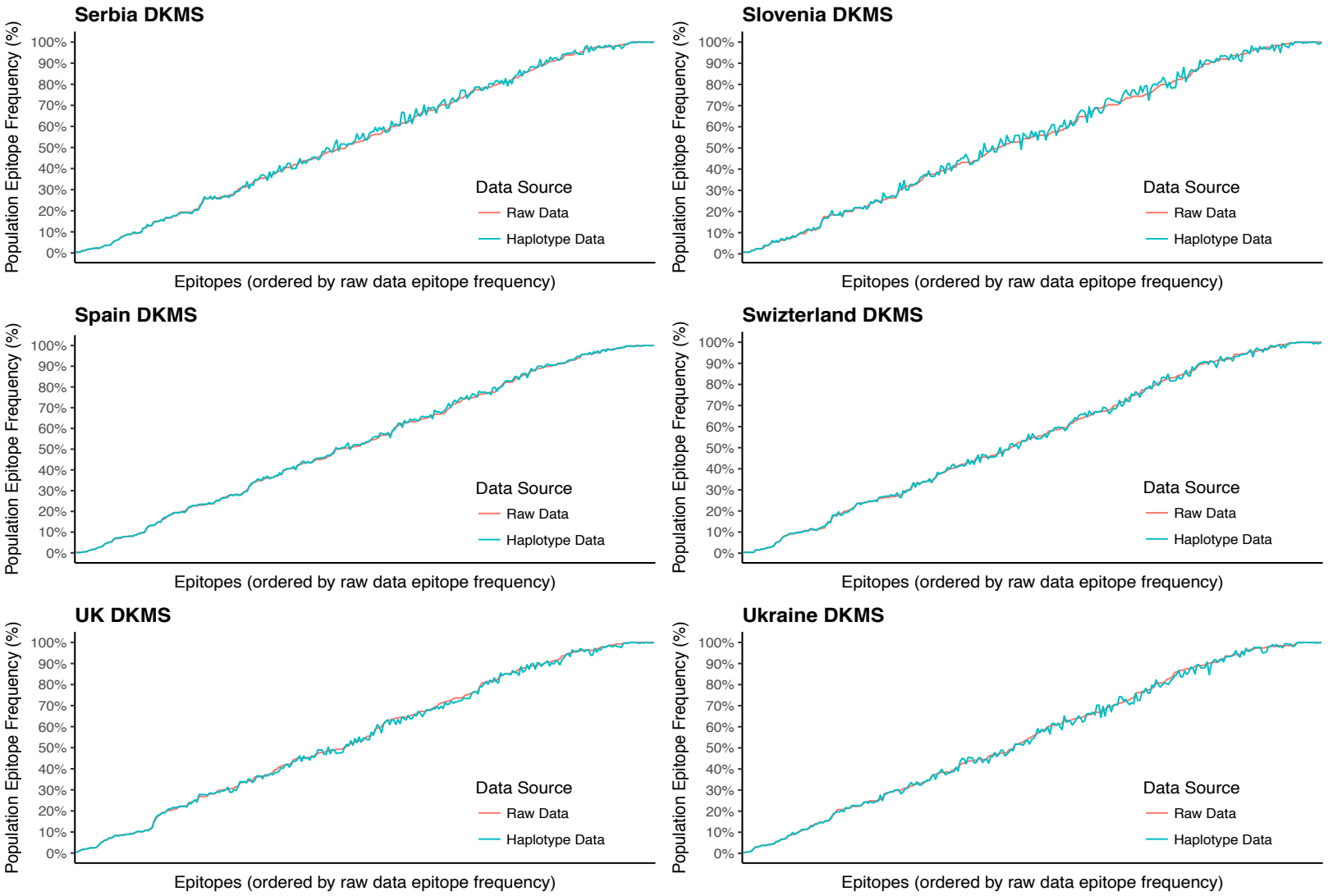

Poland DKMS
Population Epitope Frequency (%)
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Data Source
Raw Data
Haplotype Data
Epitopes (ordered by raw data epitope frequency)


Portugal Azores Terceira Island
Population Epitope Frequency (%)
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Data Source
Raw Data
Haplotype Data
Epitopes (ordered by raw data epitope frequency)


Portugal DKMS
Population Epitope Frequency (%)
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Data Source
Raw Data
Haplotype Data
Epitopes (ordered by raw data epitope frequency)


Romania DKMS
Population Epitope Frequency (%)
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Data Source
Raw Data
Haplotype Data
Epitopes (ordered by raw data epitope frequency)


Russia DKMS
Population Epitope Frequency (%)
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Data Source
Raw Data
Haplotype Data
Epitopes (ordered by raw data epitope frequency)


Serbia DKMS
Slovenia DKMS
Spain DKMS
Swizterland DKMS
UK DKMS
Ukraine DKMS
Population Epitope Frequency (%)
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Data Source
Raw Data
Haplotype Data
Epitopes (ordered by raw data epitope frequency)

Bibliography

1. Murphy K, Travers P, Walport M, Janeway C: *Janeway's immunobiology* edn 8th. New York: Garland Science; 2012.
2. Turvey SE, Broide DH: **Innate immunity**. *J Allergy Clin Immunol* 2010, **125**:S24-32.
3. Iwasaki A, Medzhitov R: **Regulation of adaptive immunity by the innate immune system**. *Science* 2010, **327**:291-295.
4. Stockwin LH, McGonagle D, Martin IG, Blair GE: **Dendritic cells: immunological sentinels with a central role in health and disease**. *Immunol Cell Biol* 2000, **78**:91-102.
5. Pancer Z, Cooper MD: **The evolution of adaptive immunity**. *Annu Rev Immunol* 2006, **24**:497-518.
6. Chaplin DD: **Overview of the immune response**. *J Allergy Clin Immunol* 2010, **125**:S3-23.
7. Bonilla FA, Oettgen HC: **Adaptive immunity**. *J Allergy Clin Immunol* 2010, **125**:S33-40.
8. Schroeder HW, Jr., Cavacini L: **Structure and function of immunoglobulins**. *J Allergy Clin Immunol* 2010, **125**:S41-52.
9. Al-Lazikani B, Lesk AM, Chothia C: **Standard conformations for the canonical structures of immunoglobulins**. *J Mol Biol* 1997, **273**:927-948.
10. Duquesnoy RJ: **A structurally based approach to determine HLA compatibility at the humoral immune level**. *Hum Immunol* 2006, **67**:847-862.
11. Lanier LL: **NK cell recognition**. *Annu Rev Immunol* 2005, **23**:225-274.
12. Caligiuri MA: **Human natural killer cells**. *Blood* 2008, **112**:461-469.
13. Vivier E, van de Pavert SA, Cooper MD, Belz GT: **The evolution of innate lymphoid cells**. *Nat Immunol* 2016, **17**:790-794.
14. Ljunggren HG, Karre K: **In search of the 'missing self': MHC molecules and NK cell recognition**. *Immunol Today* 1990, **11**:237-244.
15. Moretta L, Biassoni R, Bottino C, Mingari MC, Moretta A: **Human NK-cell receptors**. *Immunol Today* 2000, **21**:420-422.
16. Orange JS: **Formation and function of the lytic NK-cell immunological synapse**. *Nat Rev Immunol* 2008, **8**:713-725.

17. Liu WR, Kim J, Nwankwo C, Ashworth LK, Arm JP: **Genomic organization of the human leukocyte immunoglobulin-like receptors within the leukocyte receptor complex on chromosome 19q13.4**. *Immunogenetics* 2000, **51**:659-669.

18. Wende H, Colonna M, Ziegler A, Volz A: **Organization of the leukocyte receptor cluster (LRC) on human Chromosome 19q13.4**. *Mammalian Genome* 1999, **10**:154-160.

19. Bashirova AA, Martin MP, McVicar DW, Carrington M: **The killer immunoglobulin-like receptor gene cluster: tuning the genome for defense**. *Annu Rev Genomics Hum Genet* 2006, **7**:277-300.

20. Brusilovsky M, Rosental B, Shemesh A, Appel MY, Porgador A: **Human NK cell recognition of target cells in the prism of natural cytotoxicity receptors and their ligands**. *J Immunotoxicol* 2012, **9**:267-274.

21. Moretta A, Marcenaro E, Parolini S, Ferlazzo G, Moretta L: **NK cells at the interface between innate and adaptive immunity**. *Cell Death Differ* 2008, **15**:226-233.

22. Poggi A, Zocchi MR: **NK cell autoreactivity and autoimmune diseases**. *Front Immunol* 2014, **5**:27.

23. Colucci F: **The role of KIR and HLA interactions in pregnancy complications**. *Immunogenetics* 2017.

24. Kulkarni S, Martin MP, Carrington M: **The Yin and Yang of HLA and KIR in human disease**. *Semin Immunol* 2008, **20**:343-352.

25. Rezvani K, Rouce RH: **The Application of Natural Killer Cell Immunotherapy for the Treatment of Cancer**. *Front Immunol* 2015, **6**:578.

26. Simonetta F, Alvarez M, Negrin RS: **Natural Killer Cells in Graft-versus-Host-Disease after Allogeneic Hematopoietic Cell Transplantation**. *Frontiers in Immunology* 2017, **8**.

27. Littera R, Piredda G, Argiolas D, Lai S, Congeddu E, Ragatzu P, Melis M, Carta E, Michittu MB, Valentini D, et al.: **KIR and their HLA Class I ligands: Two more pieces towards completing the puzzle of chronic rejection and graft loss in kidney transplantation**. *PLoS One* 2017, **12**:e0180831.

28. Templeton AR: *Population genetics and microevolutionary theory*. Hoboken, N.J.: Wiley-Liss; 2006.

29. Single RM, Martin MP, Meyer D, Gao X, Carrington M: **Methods for assessing gene content diversity of KIR with examples from a global set of populations**. *Immunogenetics* 2008, **60**:711-725.

30. Gillespie JH: *Population genetics : a concise guide*. Baltimore, Md. ; London: The Johns Hopkins University Press; 1998.

31. Middleton D, Gonzalez F: **Immunogenetic Databases**. In *The HLA complex in Biology and Medicine: A resource book*. Edited by Mehra N: Jaypee; 2010:119-134.

32. Single R, Gourraud P-A, Lancaster A, Briggs F, F Barcellos L, Hollenbach J, Mack S, Thomson G: **Haplotype Estimation and Linkage Disequilibrium Methods Manual: Version 0.1.8**. Edited by; 2011.

33. Lancaster AK, Single RM, Solberg OD, Nelson MP, Thomson G: **PyPop update--a software pipeline for large-scale multilocus population genomics**. *Tissue Antigens* 2007, **1**:192-197.

34. Trowsdale J, Knight JC: **Major histocompatibility complex genomics and human disease**. *Annu Rev Genomics Hum Genet* 2013, **14**:301-323.

35. Blackwell JM, Jamieson SE, Burgner D: **HLA and infectious diseases**. *Clin Microbiol Rev* 2009, **22**:370-385, Table of Contents.

36. Marsh SGE, Parham P, Barber LD: *The HLA factsbook*. San Diego, California: Academic Press; 2000.

37. Klein J, Sato A: **The HLA system. First of two parts**. *N Engl J Med* 2000, **343**:702-709.

38. Bowness P, Zaccai N, Bird L, Jones EY: **HLA-B27 and disease pathogenesis: new structural and functional insights**. *Expert Rev Mol Med* 1999, **1999**:1-10.

39. Marsh SG, Albert ED, Bodmer WF, Bontrop RE, Dupont B, Erlich HA, Fernandez-Vina M, Geraghty DE, Holdsworth R, Hurley CK, et al.: **Nomenclature for factors of the HLA system, 2010**. *Tissue Antigens* 2010, **75**:291-455.

40. Listgarten J, Brumme Z, Kadie C, Xiaojiang G, Walker B, Carrington M, Goulder P, Heckerman D: **Statistical resolution of ambiguous HLA typing data**. *PLoS Comput Biol* 2008, **4**:e1000016.

41. Sanchez-Mazas A, Lemaitre JF, Currat M: **Distinct evolutionary strategies of human leucocyte antigen loci in pathogen-rich environments**. *Philos Trans R Soc Lond B Biol Sci* 2012, **367**:830-839.

42. Parham P, Norman PJ, Abi-Rached L, Guethlein LA: **Human-specific evolution of killer cell immunoglobulin-like receptor recognition of major histocompatibility complex class I molecules**. *Philos Trans R Soc Lond B Biol Sci* 2012, **367**:800-811.

43. Moffett A, Colucci F: **Co-evolution of NK receptors and HLA ligands in humans is driven by reproduction**. *Immunol Rev* 2015, **267**:283-297.

44. Gonzalez-Galarza FF, Takeshita LY, Santos EJ, Kempson F, Maia MH, Silva AL, Silva AL, Ghattaoraya GS, Alfirevic A, Jones AR, et al.: **Allele frequency net 2015 update: new features for HLA epitopes, KIR and disease and HLA adverse drug reaction associations**. *Nucleic Acids Res* 2015, **43**:20.

45. Gough SC, Simmonds MJ: **The HLA Region and Autoimmune Disease: Associations and Mechanisms of Action**. *Curr Genomics* 2007, **8**:453-465.

46. Rich SS, French LR, Sprafka JM, Clements JP, Goetz FC: **HLA-associated susceptibility to type 2 (non-insulin-dependent) diabetes mellitus: the Wadena City Health Study**. *Diabetologia* 1993, **36**:234-238.

47. Nomura S, Shouzu A, Omoto S, Matsuzaki T, Yamaoka M, Abe M, Hosokawa M, Nishikawa M, Iwasaka T, Fukuhara S: **Genetic analysis of HLA, NA and HPA typing in type 2 diabetes and ASO**. *Int J Immunogenet* 2006, **33**:117-122.

48. Yang J, Lernmark A, Uusitalo UM, Lynch KF, Veijola R, Winkler C, Larsson HE, Rewers M, She JX, Ziegler AG, et al.: **Prevalence of obesity was related to HLA-DQ in 2-4-year-old children at genetic risk for type 1 diabetes**. *Int J Obes (Lond)* 2014, **38**:1491-1496.

49. Chien YL, Wu YY, Chen CH, Gau SS, Huang YS, Chien WH, Hu FC, Chao YL: **Association of HLA-DRB1 alleles and neuropsychological function in autism**. *Psychiatr Genet* 2012, **22**:46-49.

50. Hamza TH, Zabetian CP, Tenesa A, Laederach A, Montimurro J, Yearout D, Kay DM, Doheny KF, Paschall J, Pugh E, et al.: **Common genetic variation in the HLA region is associated with late-onset sporadic Parkinson's disease**. *Nat Genet* 2010, **42**:781-785.

51. Lehmann DJ, Barnardo MCNM, Fuggle S, Quiroga I, Sutherland A, Warden DR, Barnetson L, Horton R, Beck S, Smith AD: **Replication of the association of HLA-B7 with Alzheimer's disease: a role for homozygosity?** *Journal of Neuroinflammation* 2006, **3**:33-33.

52. Ghattaoraya GS, Dundar Y, Gonzalez-Galarza FF, Maia MH, Santos EJ, da Silva AL, McCabe A, Middleton D, Alfirevic A, Dickson R, et al.: **A web resource for mining HLA associations with adverse drug reactions: HLA-ADR**. *Database (Oxford)* 2016, **2016**.

53. Marsh SG, Parham P, Dupont B, Geraghty DE, Trowsdale J, Middleton D, Vilches C, Carrington M, Witt C, Guethlein LA, et al.: **Killer-cell immunoglobulin-like receptor (KIR) nomenclature report, 2002**. *Immunogenetics* 2003, **55**:220-226.

54. Khakoo SI, Rajalingam R, Shum BP, Weidenbach K, Flodin L, Muir DG, Canavez F, Cooper SL, Valiante NM, Lanier LL, et al.: **Rapid evolution of NK cell receptor systems demonstrated by comparison of chimpanzees and humans**. *Immunity* 2000, **12**:687-698.

55. Wilson MJ, Torkar M, Haude A, Milne S, Jones T, Sheer D, Beck S, Trowsdale J: **Plasticity in the organization and sequences of human KIR/ILT gene families**. *Proceedings of the National Academy of Sciences* 2000, **97**:4778-4783.

56. Middleton D, Gonzelez F: **The extensive polymorphism of KIR genes**. *Immunology* 2010, **129**:8-19.

57. Uhrberg M, Valiante NM, Shum BP, Shilling HG, Lienert-Weidenbach K, Corliss B, Tyan D, Lanier LL, Parham P: **Human diversity in killer cell inhibitory receptor genes**. *Immunity* 1997, **7**:753-763.

58. Gonzalez-Galarza FF, Takeshita LY, Santos EJ, Kempson F, Maia MH, da Silva AL, Teles e Silva AL, Ghattaoraya GS, Alfirevic A, Jones AR, et al.: **Allele frequency net 2015 update: new features for HLA epitopes, KIR and disease and HLA adverse drug reaction associations**. *Nucleic Acids Res* 2015, **43**:D784-788.

59. Robinson J, Halliwell JA, Hayhurst JD, Flicek P, Parham P, Marsh SG: **The IPD and IMGT/HLA database: allele variant databases**. *Nucleic Acids Res* 2015, **43**:D423-431.

60. Robinson J, Halliwell JA, McWilliam H, Lopez R, Marsh SGE: **IPD—the Immuno Polymorphism Database**. *Nucleic acids research* 2013, **41**:D1234-D1240.

61. Rajalingam R, Krausa P, Shilling HG, Stein JB, Balamurugan A, McGinnis MD, Cheng NW, Mehra NK, Parham P: **Distinctive KIR and HLA diversity in a panel of north Indian Hindus**. *Immunogenetics* 2002, **53**:1009-1019.

62. Norman PJ, Carrington CV, Byng M, Maxwell LD, Curran MD, Stephens HA, Chandanayingyong D, Verity DH, Hameed K, Ramdath DD, et al.: **Natural killer cell immunoglobulin-like receptor (KIR) locus profiles in African and South Asian populations**. *Genes Immun* 2002, **3**:86-95.

63. Falco M, Moretta L, Moretta A, Bottino C: **KIR and KIR ligand polymorphism: a new area for clinical applications?** *Tissue Antigens* 2013, **82**:363-373.

64. Rajalingam R: **Overview of the killer cell immunoglobulin-like receptor system**. *Methods Mol Biol* 2012, **882**:391-414.

65. Hilton HG, Vago L, Older Aguilar AM, Moesta AK, Graef T, Abi-Rached L, Norman PJ, Guethlein LA, Fleischhauer K, Parham P: **Mutation at positively selected positions in the binding site for HLA-C shows that KIR2DL1 is a more refined but less adaptable NK cell receptor than KIR2DL3**. *J Immunol* 2012, **189**:1418-1430.

66. Rajagopalan S, Long EO: **Understanding how combinations of HLA and KIR genes influence disease**. *J Exp Med* 2005, **201**:1025-1029.

67. Rajagopalan S, Long EO: **KIR2DL4 (CD158d): An activation receptor for HLA-G**. *Front Immunol* 2012, **3**:258.

68. Rajagopalan S: **HLA-G-mediated NK cell senescence promotes vascular remodeling: implications for reproduction**. *Cell Mol Immunol* 2014, **11**:460-466.

69. Jamil KM, Khakoo SI: **KIR/HLA interactions and pathogen immunity**. *J Biomed Biotechnol* 2011, **298348**:19.

70. Phillips BL, Callaghan C: **The immunology of organ transplantation**. In *Surgery (United Kingdom)*. Edited by; 2017:333-340. vol 35.]

71. Wood KJ, Goto R: **Mechanisms of rejection: current perspectives**. *Transplantation* 2012, **93**:1-10.

72. Gaston RS, Cecka JM, Kasiske BL, Fieberg AM, Leduc R, Cosio FC, Gourishankar S, Grande J, Halloran P, Hunsicker L, et al.: **Evidence for antibody-mediated injury as a major determinant of late kidney allograft failure**. *Transplantation* 2010, **90**:68-74.

73. Sheldon S, Poulton K: **HLA Typing and Its Influence on Organ Transplantation**. In *Transplantation Immunology: Methods and Protocols*. Edited by Hornick P, Rose M: Humana Press; 2006:157-174.

74. Abecassis M, Bartlett ST, Collins AJ, Davis CL, Delmonico FL, Friedewald JJ, Hays R, Howard A, Jones E, Leichtman AB, et al.: **Kidney Transplantation as Primary Therapy for End-Stage Renal Disease: A National Kidney Foundation/Kidney Disease Outcomes Quality Initiative (NKF/KDOQI™) Conference**. *Clinical Journal of the American Society of Nephrology : CJASN* 2008, **3**:471-480.

75. Chandak P, Callaghan C: **The immunology of organ transplantation**. *Surgery (Oxford)* 2014, **32**:325-332.

76. Takemoto S, Port FK, Claas FH, Duquesnoy RJ: **HLA matching for kidney transplantation**. *Hum Immunol* 2004, **65**:1489-1505.

77. Keith DS, Vranic GM: **Approach to the Highly Sensitized Kidney Transplant Candidate**. *Clin J Am Soc Nephrol* 2016, **11**:684-693.

78. Leffell MS, Donnenberg AD, Rose NR: *Handbook of human immunology*. Boca Raton: CRC Press; 1997.

79. Persijn GG, Smits J, De Meester J, Frei U: **Three-year experience with the new Eurotransplant kidney allocation system 1996-1999**. *Transplant Proc* 2001, **33**:821-823.

80. Cecka JM: **The role of HLA in renal transplantation**. *Hum Immunol* 1997, **56**:6-16.

81. Lefranc MP, Giudicelli V, Duroux P, Jabado-Michaloud J, Folch G, Aouinti S, Carillon E, Duvergey H, Houles A, Paysan-Lafosse T, et al.: **IMGT(R), the international ImMunoGeneTics information system(R) 25 years on**. *Nucleic Acids Res* 2015, **43**:D413-422.

82. Blackwell JM, Jamieson SE, Burgner D: **HLA and infectious diseases**. *Clin Microbiol Rev* 2009, **22**:370-385.

83. Huard B, Karlsson L: **KIR expression on self-reactive CD8+ T cells is controlled by T-cell receptor engagement**. *Nature* 2000, **403**:325-328.

84. Gonzalez-Galarza FF, Christmas S, Middleton D, Jones AR: **Allele frequency net: a database and online repository for immune gene frequencies in worldwide populations**. *Nucleic Acids Res* 2011, **39**:D913-919.

85. Khakoo SI, Carrington M: **KIR and disease: a model system or system of models?** *Immunol Rev* 2006, **214**:186-201.

86. Takeshita LY, Gonzalez-Galarza FF, dos Santos EJ, Maia MH, Rahman MM, Zain SM, Middleton D, Jones AR: **A database for curating the associations between killer cell immunoglobulin-like receptors and diseases in worldwide populations**. *Database (Oxford)* 2013, **2013**:bat021.

87. Takeshita LY, Jones AR, Gonzalez-Galarza FF, Middleton D: **Allele frequencies database**. *Transfus Med Hemother* 2014, **41**:352-355.

88. Yu W, Gwinn M, Clyne M, Yesupriya A, Khoury MJ: *A navigator for human genome epidemiology*: Nat Genet. 2008 Feb;40(2):124-5.

89. R Core team: **R Core Team**. In *R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing , Vienna, Austria. ISBN 3-900051-07-0, URL* http://www.r-project.org/. Edited by; 2015:275-286. vol 55.]

90. RStudio Team -: **RStudio: Integrated Development for R**. *[Online] RStudio, Inc., Boston, MA URL* http://www/. *rstudio. com* 2016:RStudio, Inc., Boston, MA.

91. Okada H, Kuhn C, Feillet H, Bach JF: **The 'hygiene hypothesis' for autoimmune and allergic diseases: an update**. *Clin Exp Immunol* 2010, **160**:1-9.

92. Pang T: **From immune system to health systems - Challenges for health research**. *Immunol Cell Biol* 2004, **82**:149-153.

93. Mills A: **Health Care Systems in Low- and Middle-Income Countries**. *New England Journal of Medicine* 2014, **370**:552-557.

94. Hiby SE, Apps R, Chazara O, Farrell LE, Magnus P, Trogstad L, Gjessing HK, Carrington M, Moffett A: **Maternal KIR in combination with paternal HLA-C2 regulate human birth weight**. *J Immunol* 2014, **192**:5069-5073.

95. Hilton HG, Parham P: **Missing or altered self: human NK cell receptors that recognize HLA-C**. *Immunogenetics* 2017, **69**:567-579.

96. Bruce N, Pope D, Stanistreet D: **Systematic Reviews and Meta-Analysis**. In *Quantitative Methods for Health Research*. Edited by: John Wiley & Sons, Ltd; 2008:393-432.

97. Munafo MR, Flint J: **Meta-analysis of genetic association studies**. *Trends Genet* 2004, **20**:439-444.

98. Lee YH: **Meta-analysis of genetic association studies**. *Ann Lab Med* 2015, **35**:283-287.

99. Liang HL, Ma SJ, Tan HZ: **Association between killer cell immunoglobulin-like receptor (KIR) polymorphisms and systemic lupus erythematosus (SLE) in populations: A PRISMA-compliant meta-analysis**. *Medicine (Baltimore)* 2017, **96**:e6166.

100. Li X, Xia Q, Fan D, Cai G, Yang X, Wang L, Xin L, Ding N, Hu Y, Liu L, et al.: **Association between KIR gene polymorphisms and rheumatoid arthritis susceptibility: A meta-analysis**. *Hum Immunol* 2015, **76**:565-570.

101. Liu SL, Zheng AJ, Ding L: **Association between KIR gene polymorphisms and type 1 diabetes mellitus (T1DM) susceptibility: A PRISMA-compliant meta-analysis**. *Medicine (Baltimore)* 2017, **96**:e9439.

102. Fan D, Liu S, Yang T, Wu S, Wang S, Li G, Zeng Z, Duan Z, Xia G, Ye D, et al.: **Association between KIR polymorphisms and ankylosing spondylitis in populations: a meta-analysis**. *Mod Rheumatol* 2014, **24**:985-991.

103. Ghanadi K, Shayanrad B, Ahmadi SA, Shahsavar F, Eliasy H: **Colorectal cancer and the KIR genes in the human genome: A meta-analysis**. *Genom Data* 2016, **10**:118-126.

104. Gauthiez E, Habfast-Robertson I, Rueger S, Kutalik Z, Aubert V, Berg T, Cerny A, Gorgievski M, George J, Heim MH, et al.: **A systematic review and meta-analysis of HCV clearance**. *Liver Int* 2017, **37**:1431-1445.

105. Sanchez-Mazas A, Vidan-Jeras B, Nunes JM, Fischer G, Little AM, Bekmane U, Buhler S, Buus S, Claas FH, Dormoy A, et al.: **Strategies to work with HLA data in human populations for histocompatibility, clinical transplantation, epidemiology and population genetics: HLA-NET methodological recommendations**. *Int J Immunogenet* 2012, **39**:459-472.

106. Bach FH, van Rood JJ: **The Major Histocompatibility Complex — Genetics and Biology**. *New England Journal of Medicine* 1976, **295**:927-936.

107. Robinson J, Halliwell JA, McWilliam H, Lopez R, Parham P, Marsh SG: **The IMGT/HLA database**. *Nucleic Acids Res* 2013, **41**:17.

108. Kumbala D, Zhang R: *Essential concept of transplant immunology for clinical practice*: World J Transplant. 2013 Dec 24;3(4):113-118.; 2013.

109. Nowak J: **Role of HLA in hematopoietic SCT**. *Bone Marrow Transplant* 2008, **42 Suppl 2**:S71-76.

110. Bontadini A: **HLA techniques: typing and antibody detection in the laboratory of immunogenetics**. *Methods* 2012, **56**:471-476.

111. Wood KJ, Goto R: *Mechanisms of rejection: current perspectives*: Transplantation. 2012 Jan 15;93(1):1-10. doi: 10.1097/TP.0b013e31823cab44.

112. Claas FH, Doxiadis, II: **Management of the highly sensitized patient**. *Curr Opin Immunol* 2009, **21**:569-572.

113. Bray RA, Nolen JD, Larsen C, Pearson T, Newell KA, Kokko K, Guasch A, Tso P, Mendel JB, Gebel HM: **Transplanting the highly sensitized patient: The emory algorithm**. *Am J Transplant* 2006, **6**:2307-2315.

114. Haarberg KM, Tambur AR: **Detection of donor-specific antibodies in kidney transplantation**. *Br Med Bull* 2014, **110**:23-34.

115. Lachmann N, Todorova K, Schulze H, Schonemann C: **Luminex((R)) and its applications for solid organ transplantation, hematopoietic stem cell transplantation, and transfusion**. *Transfus Med Hemother* 2013, **40**:182-189.

116. Konvalinka A, Tinckam K: **Utility of HLA Antibody Testing in Kidney Transplantation**. *J Am Soc Nephrol* 2015, **26**:1489-1502.

117. Tait BD: **Detection of HLA Antibodies in Organ Transplant Recipients - Triumphs and Challenges of the Solid Phase Bead Assay**. *Front Immunol* 2016, **7**:570.

118. Sullivan HC, Gebel HM, Bray RA: **Understanding solid-phase HLA antibody assays and the value of MFI**. *Hum Immunol* 2017, **78**:471-480.

119. El-Awar N, Terasaki PI, Cai J, Deng CT, Ozawa M, Nguyen A, Lias M, Conger N: **Epitopes of HLA-A, B, C, DR, DQ, DP and MICA antigens**. *Clin Transpl* 2009:295-321.

120. Duquesnoy RJ, Marrari M, da MSLC, de MBJR, de SUAKM, da Silva AS, do Monte SJ: **16th IHIW: a website for antibody-defined HLA epitope Registry**. *Int J Immunogenet* 2013, **40**:54-59.

121. Duquesnoy RJ: **HLAMatchmaker: a molecularly based algorithm for histocompatibility determination. I. Description of the algorithm**. *Hum Immunol* 2002, **63**:339-352.

122. Duquesnoy RJ: **Clinical usefulness of HLAMatchmaker in HLA epitope matching for organ transplantation**. *Curr Opin Immunol* 2008, **20**:594-601.

123. Duquesnoy RJ: *HLA epitope based matching for transplantation*: Transpl Immunol. 2014 Apr 25. pii: S0966-3274(14)00027-6. doi: 10.1016/j.trim.2014.04.004.; 2014.

124. Duquesnoy RJ, Askar M: **HLAMatchmaker: a molecularly based algorithm for histocompatibility determination. V. Eplet matching for HLA-DR, HLA-DQ, and HLA-DP**. *Hum Immunol* 2007, **68**:12-25.

125. Anunciacao FA, Sousa LC, da Silva AS, Marroquim MS, Coelho AG, Willcox GH, de Andrade JM, Correa Bde M, Guimaraes EL, do Monte SJ: **EpViX: A cloud-based tool for epitope reactivity analysis and epitope virtual crossmatching to identify low immunologic risk donors for sensitized recipients**. *Transpl Immunol* 2015, **33**:153-158.

126. Goodman RS, Taylor CJ, O'Rourke CM, Lynch A, Bradley JA, Key T: **Utility of HLAMatchmaker and single-antigen HLA-antibody detection beads for identification of acceptable mismatches in highly sensitized patients awaiting kidney transplantation**. *Transplantation* 2006, **81**:1331-1336.

127. Claas FH, Dankers MK, Oudshoorn M, van Rood JJ, Mulder A, Roelen DL, Duquesnoy RJ, Doxiadis, II: **Differential immunogenicity of HLA mismatches in clinical transplantation**. *Transpl Immunol* 2005, **14**:187-191.

128. Dehn J, Setterholm M, Buck K, Kempenich J, Beduhn B, Gragert L, Madbouly A, Fingerson S, Maiers M: **HapLogic: A Predictive Human Leukocyte Antigen-Matching Algorithm to Enhance Rapid Identification of the Optimal Unrelated Hematopoietic Stem Cell Sources for Transplantation**. *Biol Blood Marrow Transplant* 2016, **22**:2038-2046.

129. Dubois V, Brignier A, Elsermans V, Gagne K, Kennel A, Pedron B, Picard C, Ravinet A, Varlet P, Cesbron A, et al.: **[Polymorphism in HLA and KIR genes and the impact on hematopoietic stem cell transplantation outcomes and unrelated donor selection: Guidelines from the Francophone Society of Bone Marrow Transplantation and Cellular Therapy (SFGM-TC)]**. *Bull Cancer* 2016, **103**:S243-S247.

130. Stark AE: **Hardy-Weinberg law: asymptotic approach to a generalized form**. *Science* 1976, **193**:1141-1142.

131. Bettinotti MP, Zachary AA, Leffell MS: **Clinically relevant interpretation of solid phase assays for HLA antibody**. *Curr Opin Organ Transplant* 2016, **21**:453-458.

132. Soria-Guerra RE, Nieto-Gomez R, Govea-Alonso DO, Rosales-Mendoza S: **An overview of bioinformatics tools for epitope prediction: implications on vaccine development**. *J Biomed Inform* 2015, **53**:405-414.

133. Blythe MJ, Flower DR: **Benchmarking B cell epitope prediction: underperformance of existing methods**. *Protein Sci* 2005, **14**:246-248.

134. Sela-Culang I, Kunik V, Ofran Y: **The structural basis of antibody-antigen recognition**. *Front Immunol* 2013, **4**:302.

135. Backert L, Kohlbacher O: **Immunoinformatics and epitope prediction in the age of genomic medicine**. *Genome Med* 2015, **7**:119.

136. Potocnakova L, Bhide M, Pulzova LB: **An Introduction to B-Cell Epitope Mapping and In Silico Epitope Prediction**. *J Immunol Res* 2016, **2016**:6760830.

137. Duquesnoy RJ, Marrari M, Jelenik L, Zeevi A, Claas FH, Mulder A: **Structural aspects of HLA class I epitopes reacting with human monoclonal antibodies in Ig-binding, C1q-binding and lymphocytotoxicity assays**. *Hum Immunol* 2013, **74**:1271-1279.

138. Podzamczer D, Fumero E: **The role of nevirapine in the treatment of HIV-1 disease**. *Expert Opin Pharmacother* 2001, **2**:2065-2078.

139. World Health Organization: **19th WHO Model List of Essential Medicines**. Edited by; 2015. vol 2017.]

140. Rodriguez-Arrondo F, Aguirrebengoa K, Portu J, Munoz J, Garcia MA, Goikoetxea J, Martinez E, Iribarren JA, Perez-Alvarez N, Negredo E, et al.: **Long-term effectiveness and safety outcomes in HIV-1-infected patients after a median time of 6 years on nevirapine**. *Curr HIV Res* 2009, **7**:526-532.

141. Marseille E, Kahn JG, Mmiro F, Guay L, Musoke P, Fowler MG, Jackson JB: **Cost effectiveness of single-dose nevirapine regimen for mothers and babies to decrease vertical HIV-1 transmission in sub-Saharan Africa**. *Lancet* 1999, **354**:803-809.

142. Detels R, Munoz A, McFarlane G, Kingsley LA, Margolick JB, Giorgi J, Schrager LK, Phair JP: **Effectiveness of potent antiretroviral therapy on time to AIDS and death in men with known HIV infection duration. Multicenter AIDS Cohort Study Investigators**. *JAMA* 1998, **280**:1497-1503.

143. Smerdon SJ, Jager J, Wang J, Kohlstaedt LA, Chirino AJ, Friedman JM, Rice PA, Steitz TA: **Structure of the binding site for nonnucleoside inhibitors of the reverse transcriptase of human immunodeficiency virus type 1**. *Proc Natl Acad Sci U S A* 1994, **91**:3911-3915.

144. De Clercq E: **Antiviral drugs in current clinical use**. *J Clin Virol* 2004, **30**:115-133.

145. Kharsany AB, Karim QA: **HIV Infection and AIDS in Sub-Saharan Africa: Current Status, Challenges and Opportunities**. *Open AIDS J* 2016, **10**:34-48.

146. Clotet B: **Once-daily dosing of nevirapine in HAART**. *J Antimicrob Chemother* 2008, **61**:13-16.

147. Kuznik A, Lamorde M, Hermans S, Castelnuovo B, Auerbach B, Semeere A, Sempa J, Ssennono M, Ssewankambo F, Manabe YC: **Evaluating the cost-effectiveness of combination antiretroviral therapy for the prevention of mother-to-child transmission of HIV in Uganda**. *Bull World Health Organ* 2012, **90**:595-603.

148. Kawalec P, Kryst J, Mikrut A, Pilc A: **Nevirapine-based regimens in HIV-infected antiretroviral-naive patients: systematic review and meta-analysis of randomized controlled trials**. *PLoS One* 2013, **8**:e76587.

149. Kanters S, Vitoria M, Doherty M, Socias ME, Ford N, Forrest JI, Popoff E, Bansback N, Nsanzimana S, Thorlund K, et al.: **Comparative efficacy and safety of first-line antiretroviral therapy for the treatment of HIV infection: a systematic review and network meta-analysis**. *Lancet HIV* 2016, **3**:e510-e520.

150. Popovic M, Shenton JM, Chen J, Baban A, Tharmanathan T, Mannargudi B, Abdulla D, Uetrecht JP: **Nevirapine hypersensitivity**. *Handb Exp Pharmacol* 2010:437-451.

151. Pavlos R, Mallal S, Ostrov D, Buus S, Metushi I, Peters B, Phillips E: **T cell-mediated hypersensitivity reactions to drugs**. *Annu Rev Med* 2015, **66**:439-454.

152. Srivastava A, Maggs JL, Antoine DJ, Williams DP, Smith DA, Park BK: **Role of reactive metabolites in drug-induced hepatotoxicity**. *Handb Exp Pharmacol* 2010:165-194.

153. Padmanabhan S, White KD, Gaudieri S, Phillips EJ: **Chapter 21 – HLA and the Pharmacogenomics of Drug Hypersensitivity**. *Handbook of Pharmacogenomics and Stratified Medicine* 2014:437-465.

154. Chung W-h, Wang C-w, Dao R-l: **Severe cutaneous adverse drug reactions**. 2016:758-766.

155. Hausmann O, Schnyder B, Pichler WJ: **Drug hypersensitivity reactions involving skin**. *Handb Exp Pharmacol* 2010:29-55.

156. Ghattaoraya GS, Middleton D, Santos EJ, Dickson R, Jones AR, Alfirevic A: **Human leucocyte antigen-adverse drug reaction associations: from a perspective of ethnicity**. *Int J Immunogenet* 2017, **44**:7-26.

157. Cornejo Castro EM, Carr DF, Jorgensen AL, Alfirevic A, Pirmohamed M: **HLA-allelotype associations with nevirapine-induced hypersensitivity reactions and hepatotoxicity: a systematic review of the literature and meta-analysis**. *Pharmacogenet Genomics* 2015, **25**:186-198.

158. Carr DF, Chaponda M, Jorgensen AL, Castro EC, van Oosterhout JJ, Khoo SH, Lalloo DG, Heyderman RS, Alfirevic A, Pirmohamed M: **Association of human leukocyte antigen alleles and nevirapine hypersensitivity in a Malawian HIV-infected population**. *Clin Infect Dis* 2013, **56**:1330-1339.

159. Carr DF, Bourgeois S, Chaponda M, Takeshita LY, Morris AP, Castro EM, Alfirevic A, Jones AR, Rigden DJ, Haldenby S, et al.: **Genome-wide association study**

of nevirapine hypersensitivity in a sub-Saharan African HIV-infected population. *J Antimicrob Chemother* 2017, **72**:1152-1162.

160. Illing PT, Purcell AW, McCluskey J: **The role of HLA genes in pharmacogenomics: unravelling HLA associated adverse drug reactions**. *Immunogenetics* 2017, **69**:617-630.

161. Illing PT, Vivian JP, Dudek NL, Kostenko L, Chen Z, Bharadwaj M, Miles JJ, Kjer-Nielsen L, Gras S, Williamson NA, et al.: **Immune self-reactivity triggered by drug-modified HLA-peptide repertoire**. *Nature* 2012, **486**:554-558.

162. Ostrov DA, Grant BJ, Pompeu YA, Sidney J, Harndahl M, Southwood S, Oseroff C, Lu S, Jakoncic J, de Oliveira CA, et al.: **Drug hypersensitivity caused by alteration of the MHC-presented self-peptide repertoire**. *Proc Natl Acad Sci U S A* 2012, **109**:9959-9964.

163. Chen J, Mannargudi BM, Xu L, Uetrecht J: **Demonstration of the metabolic pathway responsible for nevirapine-induced skin rash**. *Chem Res Toxicol* 2008, **21**:1862-1870.

164. The UniProt C: **UniProt: the universal protein knowledgebase**. *Nucleic Acids Res* 2017, **45**:D158-D169.

165. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: **The Protein Data Bank**. *Nucleic Acids Res* 2000, **28**:235-242.

166. Fan QR, Wiley DC: **Structure of human histocompatibility leukocyte antigen (HLA)-Cw4, a ligand for the KIR2D natural killer cell inhibitory receptor**. *J Exp Med* 1999, **190**:113-123.

167. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL: **BLAST+: architecture and applications**. *BMC Bioinformatics* 2009, **10**:421.

168. Song Y, DiMaio F, Wang RY, Kim D, Miles C, Brunette T, Thompson J, Baker D: **High-resolution comparative modeling with RosettaCM**. *Structure* 2013, **21**:1735-1742.

169. Soding J: **Protein homology detection by HMM-HMM comparison**. *Bioinformatics* 2005, **21**:951-960.

170. Nivon LG, Moretti R, Baker D: **A Pareto-optimal refinement method for protein design scaffolds**. *PLoS One* 2013, **8**:e59004.

171. Morris GM, Huey R, Lindstrom W, Sanner MF, Belew RK, Goodsell DS, Olson AJ: **AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility**. *J Comput Chem* 2009, **30**:2785-2791.

172. Schrodinger, LLC: **The PyMOL Molecular Graphics System, Version 1.8**. Edited by; 2015.

173. Trott O, Olson AJ: **AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading**. *J Comput Chem* 2010, **31**:455-461.

174. Wang Z, Sun H, Yao X, Li D, Xu L, Li Y, Tian S, Hou T: **Comprehensive evaluation of ten docking programs on a diverse set of protein-ligand complexes: the prediction accuracy of sampling power and scoring power**. *Phys Chem Chem Phys* 2016, **18**:12964-12975.

175. Older Aguilar AM, Guethlein LA, Adams EJ, Abi-Rached L, Moesta AK, Parham P: **Coevolution of killer cell Ig-like receptors with HLA-C to become the major variable regulators of human NK cells**. *J Immunol* 2010, **185**:4238-4251.

176. Winter CC, Gumperz JE, Parham P, Long EO, Wagtmann N: **Direct binding and functional transfer of NK cell inhibitory receptors reveal novel patterns of HLA-C allotype recognition**. *J Immunol* 1998, **161**:571-577.

177. Ramsbottom KA, Carr D, Jones AR, Rigden DJ: **Critical assessment of approaches for molecular docking to elucidate associations of HLA alleles with Adverse Drug Reactions**. *bioRxiv* 2018.